Faculty of Science, Technology, Engineering and Mathematics
M140 Introducing statistics

Unit 9 Comparing schools

Contents

Introduction

In Books 3 and 4 so far, we have looked at questions related to education. In Unit 6 we looked at the question: *How often do pupils truant?* The theme of Unit 7 was the question: *What factors affect a child's reading ability?* The focus of Unit 8 was: *What is the best way of teaching how to read?* We have looked at various statistical techniques to help answer these questions. For example, in Unit 6 you used the sign test to investigate truancy rates. And in Unit 7 you used a z-test to test hypotheses about the reading scores of seven- and eight-year-old British children. In this unit, we will consider the following question:

How good is a school?

This is a question that is of interest to politicians as well as parents and teachers.

In the UK general election of 2010, the three main political parties all made reference to the quality of schools in their manifestos. Quotes include:

'Raise standards in schools' – Conservatives

'Every school a good school' – Labour

'Ensure that every neighbourhood is served by an excellent local school or college' – Liberal Democrats

Charles Dickens's character Nicholas Nickleby tries to improve the standard of Dotheboys Hall School

In Section 1 we will consider how to measure the quality of a school. We will narrow the definition of school quality to something that can be objectively measured, and present some data. Section 2 introduces the correlation coefficient – a measure of the strength of the relationship between two variables. In Section 3 you will learn some more about the properties of the correlation coefficient. In Section 4 you will learn how to construct intervals for means that complement one-sample z-tests and two-sample z-tests – intervals that are known as *confidence intervals*. In Section 5 you will learn about making two types of intervals for predictions from a regression line, intervals that are known as *confidence intervals for the mean response* and *prediction intervals*. Finally, Section 6 will return to the measurement of the quality of a school.

Note that you will be guided to the Computer Book in Subsections 2.4 and 5.3. Similarly to previous units, it is better if you do the work at these points in the text, although you can leave it until later if you prefer.

1 Measuring school quality

The theme of this unit is the question: *How good is a school?* The problem with this question is one of definition. What makes a school good? How can we measure a school's quality? So, to make progress, it is necessary to clarify what we mean by a 'good school'.

1.1 Clarifying the question

In order to judge the quality of a school it is necessary to know what makes some good and others not-so-good.

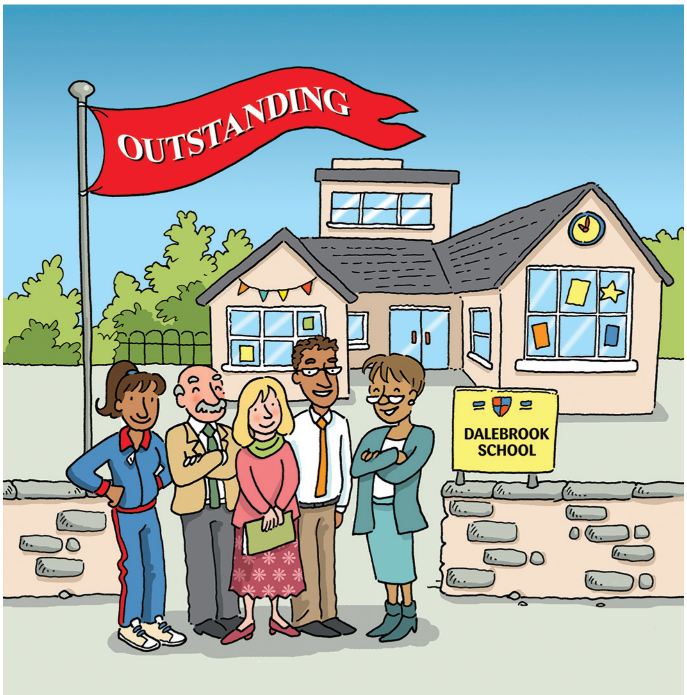
Activity 1 Thinking about the definition

Spend a few minutes thinking about what makes a school good, and list some of your thoughts.

For this unit we will just consider one measure of school quality, the academic achievement of its students.

Activity 2 Measuring academic achievement

Think about the ways in which the academic ability of a student can be measured.



In this unit we will focus on measuring the quality of secondary schools in England. In particular we will focus on the academic achievement of students at the end of one phase of their education – the end of what is known as Key Stage 4.

Key Stages

In England, many state schools must follow a curriculum laid down by the government – the 'National Curriculum'. The National Curriculum is broken down into a number of 'Key Stages', each stage corresponding to a number of year groups at school.

Key Stages

	Ages	School years
Key Stage 1	5 to 7	1 and 2
Key Stage 2	7 to 11	3 to 6
Key Stage 3	11 to 14	7 to 9
Key Stage 4	14 to 16	10 and 11

(Source: Gov.uk (2013) 'The national curriculum')

In Wales and Northern Ireland, the curriculum is also broken up into Key Stages. Although the details vary, in all three countries the end of Key Stage 2 corresponds to the end of primary education and the end of Key Stage 4 corresponds to the end of compulsory education at age 16.

The academic achievement of students when they finish Key Stage 4 can be measured by looking at how many, and which, national qualifications they are awarded at that point in their education. In particular, this can be done by considering one set of qualifications that students in England generally finish at the end of Key Stage 4 – GCSEs.

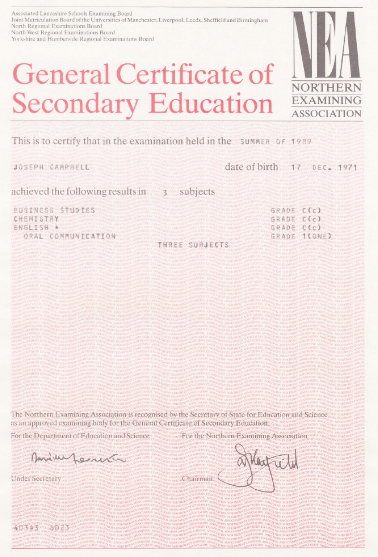
GCSE – General Certificate of Secondary Education

GCSEs (General Certificates of Secondary Education) are qualifications that are generally awarded to students at the end of Key Stage 4 in England, Wales and Northern Ireland. (Students in Scotland usually take Standard Grades instead of GCSEs.) The qualification was first awarded in 1988 and replaced GCE 'O'-Levels and CSEs.

GCSE grades of pass (in order from highest to lowest) are as follows: A*, A, B, C, D, E, F and G.

Grades A* to C correspond to Level 2 qualifications on the National Qualifications Framework for England and Northern Ireland and the Credit and Qualifications Framework for Wales – the same as Credit Standard Grades in Scotland and, at the time the qualification was introduced, the same as GCE 'O' levels at grades A to C and CSEs at grade 1.

Grades D to G correspond to Level 1 qualifications on the same frameworks – the same as General Standard Grades in Scotland and, at the time the qualifications were introduced, the same as CSEs at grades 2 to 5.



Associated Lancashire Schools Examining Board
Joint Matriculation Board of the Universities of Manchester, Liverpool, Leeds, Sheffield and Birmingham
North Regional Examinations Board
North West Regional Examinations Board
Yorkshire and Humberside Regional Examinations Board

General Certificate of Secondary Education

NEA
NORTHERN EXAMINING ASSOCIATION

This is to certify that in the examination held in the SUMMER OF 1989

JOSEPH CAMPBELL date of birth 17 DEC. 1971

achieved the following results in 3 subjects

BUSINESS STUDIES	GRADE C(c)
CHEMISTRY	GRADE C(c)
ENGLISH *	GRADE C(c)
ORAL COMMUNICATION	GRADE 1(ONE)

THREE SUBJECTS

The Northern Examining Association is recognised by the Secretary of State for Education and Science as an approved examining body for the General Certificate of Secondary Education.

For the Department of Education and Science
Under Secretary

For the Northern Examining Association
Chairman

40343 6023

As students at a secondary school generally take a variety of GCSEs, both in terms of numbers and subjects, it is necessary to combine the students' individual results in some way to get a measure of overall academic achievement by students at the school. One such measure that is commonly used at the time of writing is as follows.

GCSE headline figure, P_{KS4}

A measure of a secondary school's quality will be taken to be the percentage of students ending Key Stage 4 who achieve at least five grade A* to C GCSEs, including English and Mathematics.

This percentage will be denoted as $\boldsymbol{P}_{\textbf{KS4}}$ and referred to as a school's **GCSE headline figure**.

So, using P_{KS4} it is possible to measure a secondary school's quality. (Assuming that the information is available. It is not published for all English secondary schools.) This provides a numerical answer to the question: *How good is a school?* However, there is a problem: the numerical answer needs to be put into context.

Activity 3 Is a local secondary school good enough?

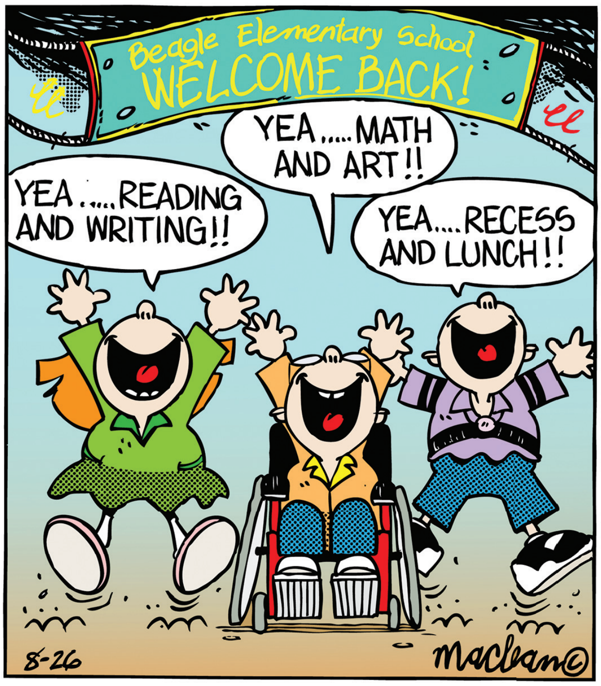
Suppose an English secondary school has a GCSE headline figure of 64%. Is this school good?

The threshold for a 'good' secondary school could be an absolute one. For example, the government could decide that the P_{KS4} in any good school should be at least 50%. But the threshold can be also set in a relative sense. For example, a secondary school might be classified as 'good' if its P_{KS4} equals or exceeds the median. To decide such relative thresholds, it is important to have a feel for the range of P_{KS4} that usually occurs. For this we need some data.

A minimum standard for English secondary schools

In a White Paper for 2010, the UK government set out a 'floor' standard for English secondary schools – that is, a minimum standard that every secondary school is expected to achieve. The White Paper (paragraph 6.26) states that:

> a school will be below the floor if fewer than 35 per cent of pupils achieve the 'basics' standard of 5 A*–C grade GCSEs including English and Mathematics, and fewer pupils make good progress between key stage two and key stage four than the national average.
>
> (Source: Department for Education (2010) *The Importance of Teaching: Schools White Paper*, London: The Stationery Office)



1.2 The data to be used

In Subsection 1.1 the quality of a secondary school was defined to be P_{KS4}, the percentage of students ending Key Stage 4 who achieve at least five grade A* to C GCSEs, including English and Mathematics. It was noted that it is important to be able to put a value of P_{KS4} into context. We will do this by considering how P_{KS4} varies for a particular group of secondary schools.

The sample we will use is a random sample of 100 English state secondary schools. Each of the schools chosen had at least 100 students ending Key Stage 4 in 2011. Furthermore, only 'non-selective' schools were chosen – that is, schools where the admissions policy does not include selection on the basis of ability.

At the time of writing, data about all state secondary schools in England are released annually by the Department for Education. These data include P_{KS4}, and so information is known about the entire population for 2011. However, some of the questions we will look at in this unit will be relevant to other years. This might include future years. So, in this sense, the values of P_{KS4} are not known for the whole population.

Histograms were introduced in Subsection 1.5 of the Computer Book.

A histogram of P_{KS4} for the 100 schools is shown in Figure 1.

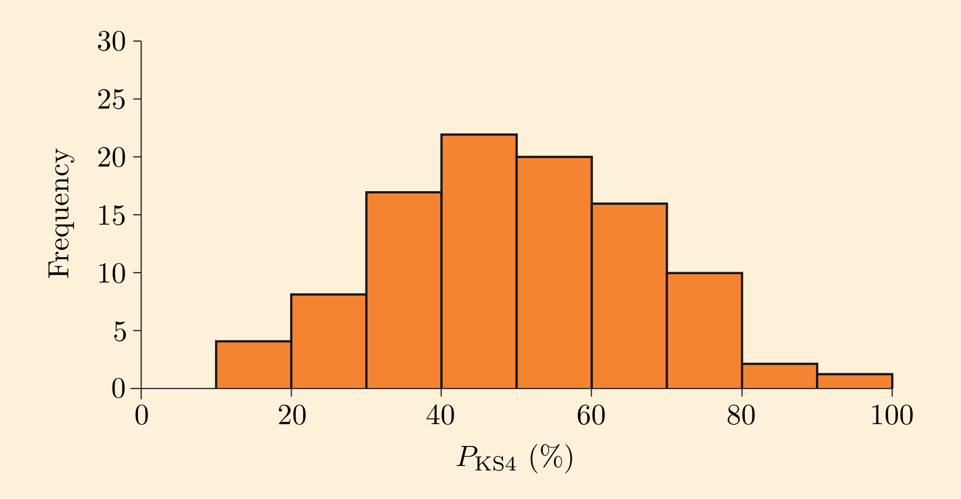


Figure 1 Histogram of P_{KS4} in 100 schools

Activity 4 Describing the distribution

Use the histogram given in Figure 1 to describe the distribution of P_{KS4} in our sample of 100 secondary schools.

Activity 5 Determining 'good enough' schools

The heights of each of the bars in Figure 1 are given in Table 1.

Table 1 P_{KS4} in 100 schools

P_{KS4}	Number of schools
0% to just under 10%	0
10% to just under 20%	4
20% to just under 30%	8
30% to just under 40%	17
40% to just under 50%	22
50% to just under 60%	20
60% to just under 70%	16
70% to just under 80%	10
80% to just under 90%	2
90% to just under 100%	1

Using this information, how many schools in our sample would be deemed to be not good enough if the following criteria for 'good' were used?

(a) $P_{\text{KS4}} \geq 50\%$

(b) $P_{\text{KS4}} \geq 30\%$

(c) $P_{\text{KS4}} \geq 90\%$

So, as Figure 1 shows, the GCSE headline figure varies considerably in our sample of 100 secondary schools. Such information allows the 'better' schools to be identified. What then? Individual parents might use this information when deciding the school they wish their child to attend. Those in charge of education, locally or nationally, might also use this information to decide which schools are not performing well enough and hence need to improve. But it does not suggest how such schools might achieve such improvement. For this a second question is important: *What factors influence the quality of a school?*

Some potential factors just split schools into two groups. For example:

- selective or non-selective
- single sex or co-educational
- in London or in the rest of England.

In Unit 7, you met a method for investigating such potential factors – the two-sample z-test. Recall that the two-sample z-test works by comparing the mean of a variable for the two groups. If the difference is big enough compared with the estimated standard error, there is evidence that the population means differ.

Activity 6 Does the type of school matter?

In the state sector, secondary schools are not all managed in the same way. One type of school is a 'community school'. So, a question that can be posed is:

> *On average, is the GCSE headline figure (P_{KS4}) the same in community schools as in other schools?*

For schools in our sample, the following summary statistics were obtained.

Table 2 Summary statistics for P_{KS4} by type of school

	Sample size	Sample mean (%)	Sample standard deviation (%)
Community school	43	49.8	13.55
Other school	57	50.7	19.61

The procedure for carrying out a two-sample z-test was detailed in Section 6 of Unit 7.

Carry out a two-sided two-sample z-test to investigate whether the population mean of P_{KS4} in community schools and the population mean of P_{KS4} in other schools are equal.

However, as you may already have realised, not all potential factors neatly split schools into two groups. For example, one possible factor that might influence a school's GCSE headline figure is the academic ability of students before they join the school. This can be measured, if imperfectly, by considering the performance in national tests of students when they ended Key Stage 2.

National Curriculum tests – SATs

Towards the end of Key Stage 2, children in English schools take National Curriculum tests, often referred to as 'SATs', in English, Mathematics and Science. In each test, a child is recorded as being at Level 5, Level 4, Level 3, Level 2 or below Level 2. These levels are then converted into points using the following scale.

Level 5: 33 points, Level 4: 27 points, Level 3: 21 points,
Level 2 or below: 15 points.

(Source: Department for Education (2011) 'Test and examination point scores used in the 2011 school and college performance tables')

So one measure of the ability of students when they join a secondary school is as follows.

P_{KS2} – the average points score (APS) at the end of Key Stage 2

For the 100 schools in our sample, the value of P_{KS2} for the students just finishing Key Stage 4 in 2011 ranges from 24.2 to 30.5. A histogram of P_{KS2} (Figure 2) shows that the values do not fall neatly into two groups. So, the two-sample z-test cannot be used to investigate whether P_{KS2} influences P_{KS4}. A different approach is needed.

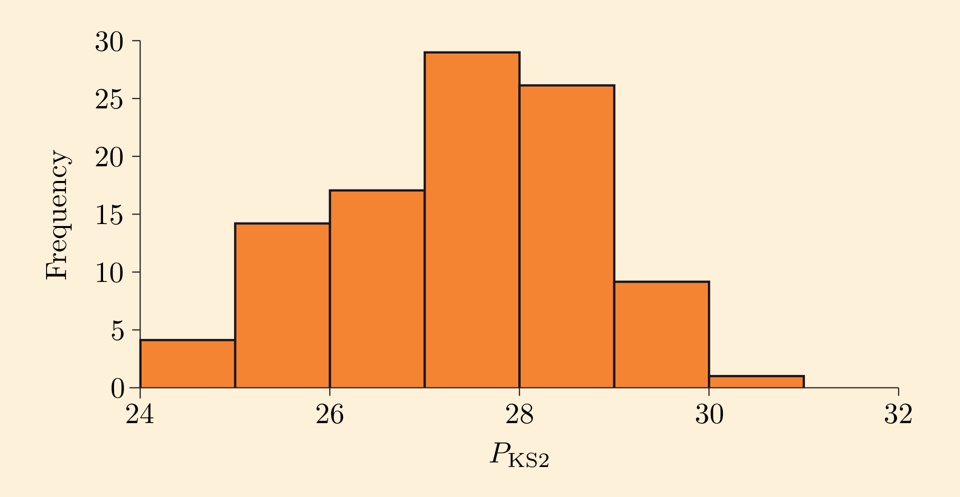


Figure 2 Histogram of P_{KS2} at intake for students at the end of Key Stage 4 in 100 schools

The variables P_{KS4} and P_{KS2} are an example of linked data. A value of P_{KS4} and P_{KS2} is available for each school in our sample. In Unit 5 you explored relationships in linked data by fitting lines. These lines summarise the relationship between two variables. However, the lines do not provide information about how strong the relationship is – that is, how accurately one variable can be predicted from knowledge about the value of the other variable. In the next section a measure of the strength of relationships is introduced – the correlation coefficient.

Linked data were introduced in Subsection 1.2 of Unit 5.

Exercises on Section 1

Exercise 1 Use a two-sample z-test?

Below, four possible factors that might affect the quality of a secondary school are listed (where the quality of a secondary school is assumed to be measured using P_{KS4}). Which of the following factors could be tested using a z-test?

(a) Whether the school has a sixth form.

(b) The size of the school (as measured by the number of students finishing Key Stage 4).

(c) Whether the school allows students to take some GCSEs a year early.

(d) The proportion of students eligible for free school meals.

2 The correlation coefficient

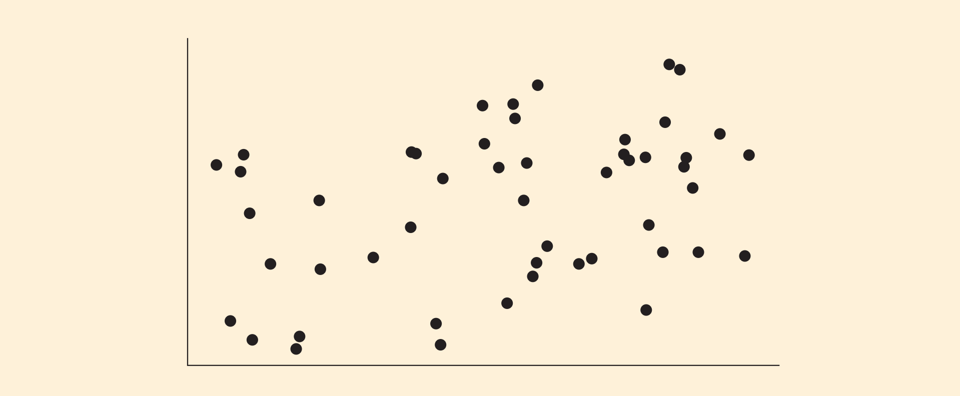
As briefly mentioned at the end of Section 1, the correlation coefficient is a measure of the strength of a relationship between two linked variables. Before going on to find out how the correlation coefficient is calculated by hand in Subsection 2.2, we first focus on what range of values a correlation coefficient can take. (Note that in Subsection 2.4 you will be referred to the Computer Book to learn how to calculate the correlation coefficient using Minitab.)

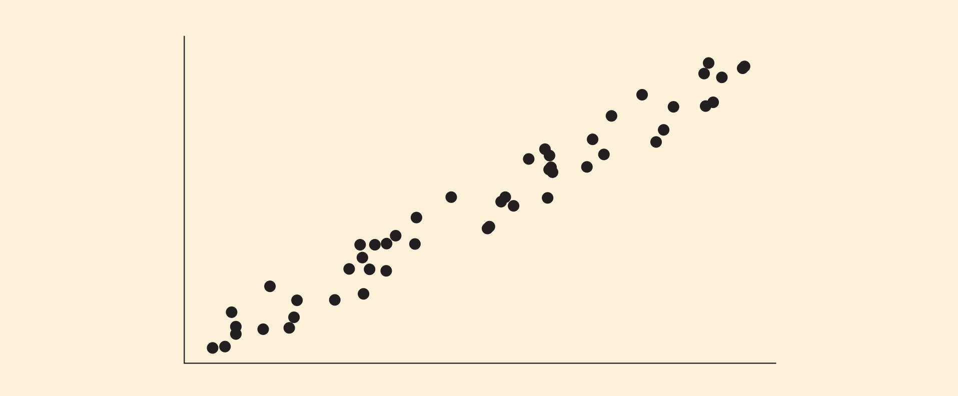
2.1 Introducing the correlation coefficient

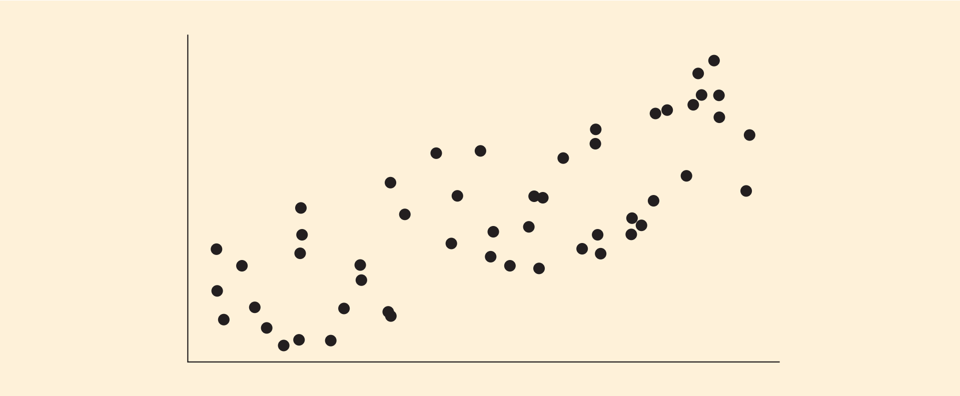
In Subsection 2.3 of Unit 5, the concept of the strength of relationships was introduced. In that unit, strength was judged subjectively from a scatterplot. Strong relationships are those where the points lie close to a line. Conversely, weak relationships are those where the points on the scatterplot only loosely follow a line.

Activity 7 Strong or weak?

For each of the scatterplots shown below, discuss whether the relationship between the two variables is strong or weak.





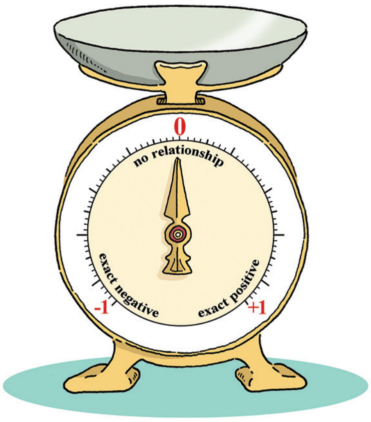


It is difficult to say how strong a relationship is just by looking at a scatterplot. The scatterplot only gives us a subjective impression of the strength of a relationship. What is needed is a single numerical quantity to summarise the strength of relationship between two variables – a *correlation coefficient*.

Correlation coefficient

A **correlation coefficient** is a number which summarises the strength of relationship between two variables.

A correlation coefficient always takes a value between -1 and $+1$. When there is an exact positive linear relationship between two variables, the value of the correlation coefficient is equal to $+1$, its maximum value. When there is an exact negative linear relationship between two variables, the value of the correlation coefficient is -1, its minimum value.



More than one correlation coefficient has been invented by statisticians – each measuring the strength of a relationship in a particular way. The coefficient we use in this module is sometimes called the 'Pearson product–moment correlation coefficient', to distinguish it from the other correlation coefficients.

Example 1 Relationships with a correlation coefficient of $+1$

In Figure 3 two scatterplots are shown. In each scatterplot the correlation coefficient is $+1$ because both depict an exact positive linear relationship. Note that the slope of the relationship is not important.

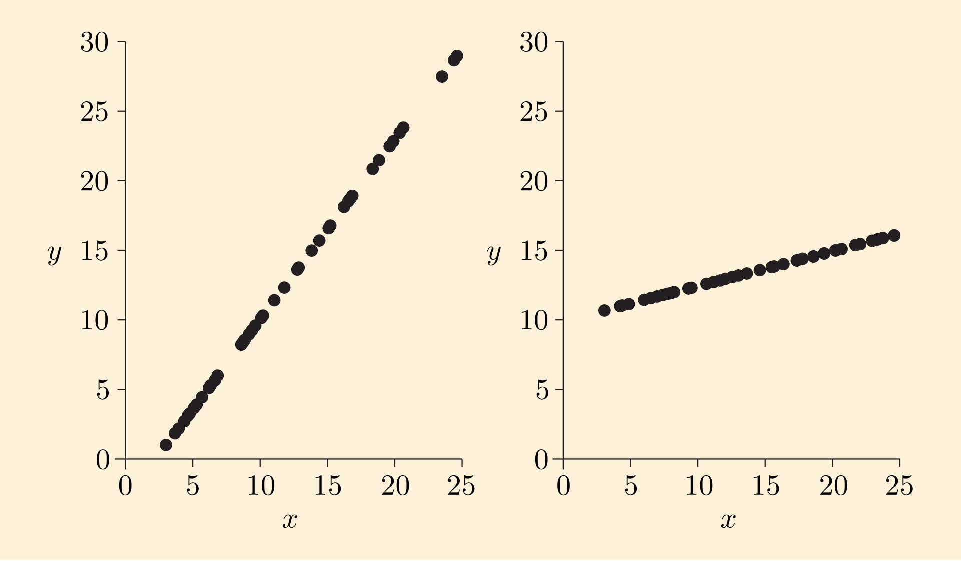


Figure 3 Scatterplots of two sets of artificial data

Example 2 Relationships with a correlation coefficient of -1

Two examples of relationships which have a correlation coefficient of -1 are shown in Figure 4. Notice that in both cases the points all lie exactly in a line and that the slope of the line is negative (that is, the line goes down from left to right).

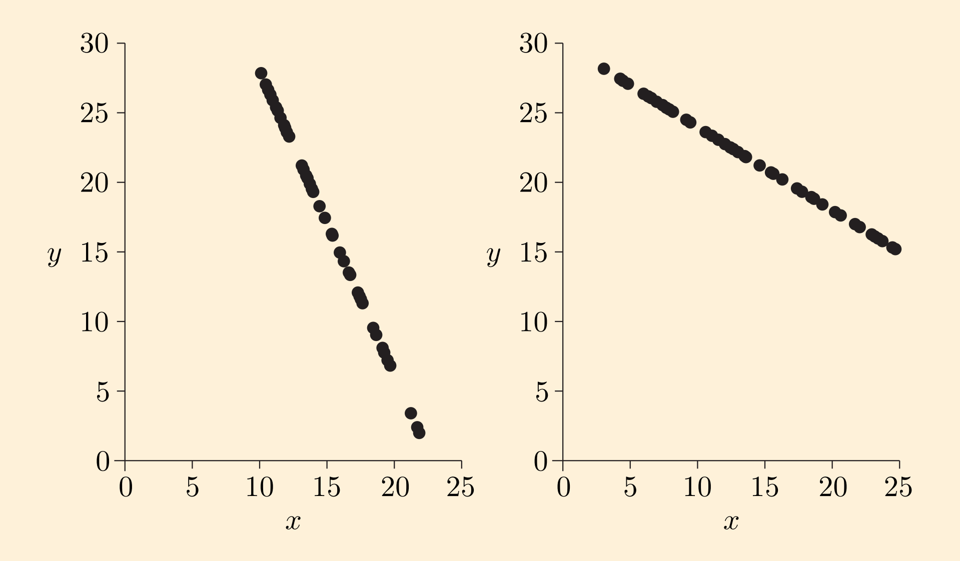
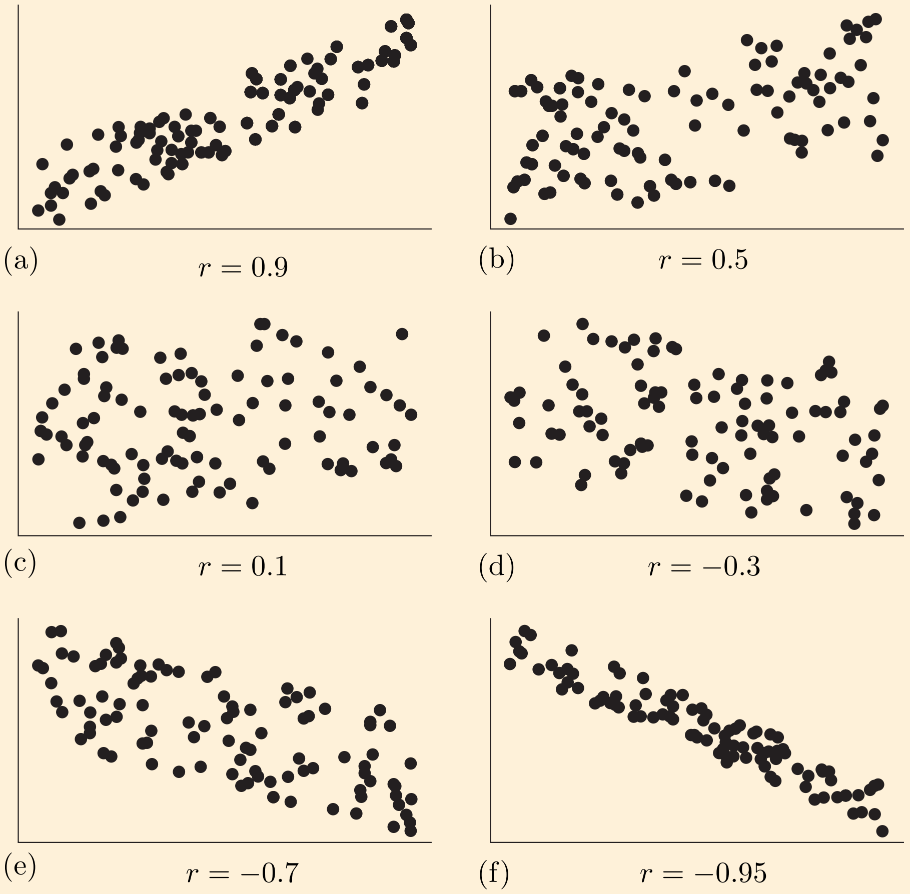


Figure 4 Scatterplots of two more sets of artificial data

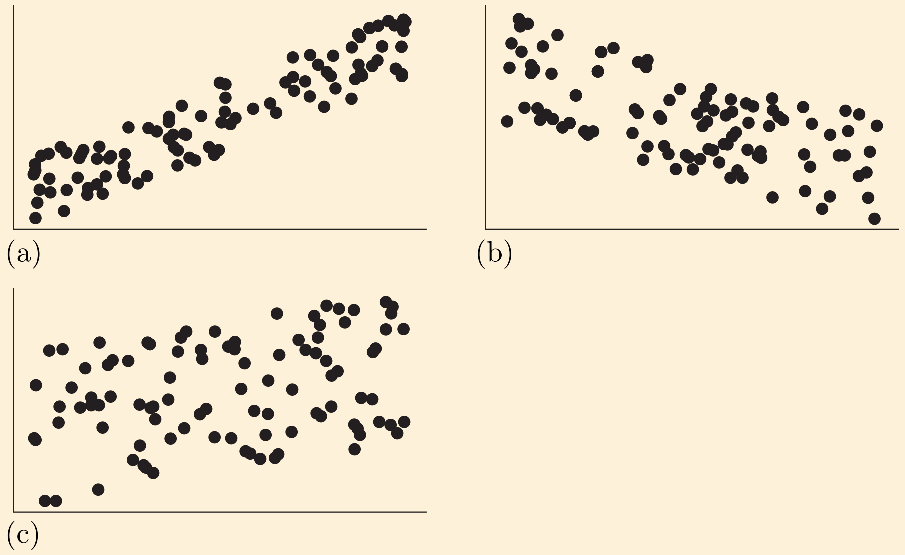
For any relationship that is not exactly linear, the correlation coefficient lies somewhere between $+1$ and -1. A positive relationship has a positive correlation coefficient, and a negative relationship has a negative coefficient. If there is no relationship between the two variables, then the correlation coefficient is equal to, or very close to, zero.

In Figure 2.1 and for the remainder of this module, the symbol r is used for the correlation coefficient. If you look at Figure 2.1 now, you will start to get a feeling for what is implied by various values of r. In (a), the points all lie quite close to an imaginary straight line sloping upwards (from left to right), and the correlation coefficient is $+0.9$. If we drew an area to include all the points, it would be long and thin. In (b) there is also a positive relationship between the two variables, but it is much less pronounced and the coefficient is only $+0.5$. An area including these points would be much fatter than that for (a). In (c), we cannot really see any relationship between the two variables; if we drew an area to include these points, it would more or less cover the whole plot. This is confirmed by the small value of r, which is $+0.1$. The plots in (d), (e) and (f) all show negative correlations.



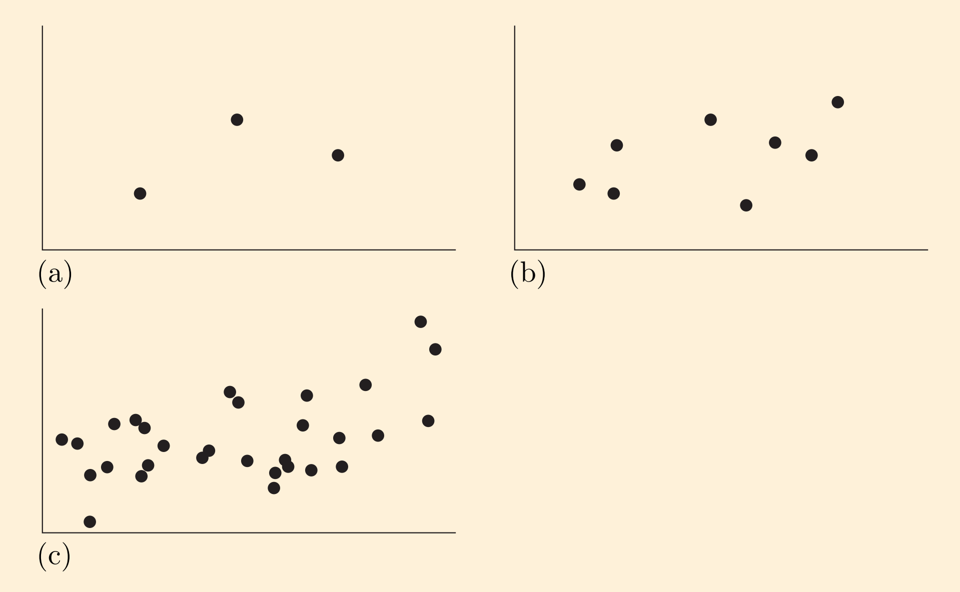
Activity 8 Estimating correlations

Look at Figure 8 and guess the correlation coefficient in each of the cases shown in (a), (b) and (c). You need not be surprised if you do not get very close to the true values, but you should at least get the sign right and probably be within about 0.2 or 0.3.



In the special case that the two data points are exactly horizontally or vertically aligned, r is undefined.

One point worth mentioning is that a correlation coefficient is much more meaningful if it is calculated from a large number of data points. If there are only two data points, then a straight line can always be drawn between them. Thus r will equal $+1$ (if the line slopes upwards) or -1 (if it slopes downwards). Neither of these values is appropriate unless there is a precise mathematical relationship between the variables. The correlation coefficient takes the value 0.5 in each of the cases (a), (b) and (c) in Figure 2.1, but the relationship is most convincing in the scatterplot in (c) because it is based on more data.



You have now covered the material related to Screencast 1 for Unit 9 (see the M140 website).

2.2 Calculating the correlation coefficient by hand

In this subsection the calculation of the correlation coefficient by hand will be demonstrated. This is done in Examples 3 to 5, where the correlation coefficient is calculated for the relationship between (i) the performance of groups of students at Key Stage 2 and (ii) the subsequent performance of broadly the same groups of students at Key Stage 4.

The following is the formula for the correlation coefficient:

$$\text{Correlation} = \frac{\sum(x-\overline{x})(y-\overline{y})}{\sqrt{\sum(x-\overline{x})^2 \times \sum(y-\overline{y})^2}}$$

If you compare this formula with that for the slope of the least squares regression line (see Subsection 4.2 of Unit 5), you will notice that some of the elements that make up the formulas are the same. This similarity means that the calculation of the correlation coefficient has a lot in common with the calculation of the least squares regression line.

Like the calculation of the slope of the least squares regression line, the first step includes calculating the sum of all the x-values ($\sum x$), the sum of all the y-values ($\sum y$), the sum of the squares of all the x-values ($\sum x^2$), and the sum of the products of the x- and y-values $(\sum xy)$. It also includes the calculation of an extra sum, the sum of the squares of all the y-values $(\sum y^2)$. In summary, step 1 is the calculation of

$$\sum x, \quad \sum y, \quad \sum x^2, \quad \sum y^2 \quad \text{and} \quad \sum xy.$$

Example 3 Calculating a correlation coefficient – step 1

The data we will use relate to students in Wales. In 2011, school-level data for Welsh schools were not made publicly available. Instead the data were just grouped by Welsh Assembly constituency. For the purpose of this calculation we look at eight of these constituencies, taking performances at Key Stage 4 in 2011 as the y-values. Students who ended Key Stage 4 in 2011 would have ended Key Stage 2 in 2006 – the constituencies' performances at Key Stage 2 in 2006 will be the x-values.

More specifically, the measure of performance at Key Stage 4 (y) will be the percentage of students achieving the following benchmark (similar to the GCSE headline figure, P_{KS4}) – the equivalent of five grade A* to C GCSEs including GCSE Mathematics and GCSE English or GCSE Welsh first language. The measure at Key Stage 2 (x) will be the percentage of students who attained the following benchmark – at least Level 4 in specified subjects. We will calculate the correlation coefficient for the relationship between these two measures.

The data are as follows:

x	78.9	75.8	77.3	74.2	78.1	72.8	77.6	77.9
y	56.7	53.1	56.1	55.9	54.1	48.6	59.4	54.0

The sums of all the x-values and of all the y-values are as follows:

$$\sum x = 78.9 + 75.8 + \cdots + 77.9 = 612.6,$$

$$\sum y = 56.7 + 53.1 + \cdots + 54.0 = 437.9.$$

We also require the sum of squares of the x-values, the sum of squares of the y-values and the sum of products of the x- and y-values. You should be able to find these three sums on your calculator without writing down each square (or product) separately.

$$\begin{aligned}\sum x^2 &= 78.9^2 + 75.8^2 + \cdots + 77.9^2 \\ &= 46\,941.4,\end{aligned}$$

$$\begin{aligned}\sum y^2 &= 56.7^2 + 53.1^2 + \cdots + 54.0^2 \\ &= 24\,039.65,\end{aligned}$$

$$\begin{aligned}\sum xy &= 78.9 \times 56.7 + 75.8 \times 53.1 + \cdots + 77.9 \times 54.0 \\ &= 33\,562.25.\end{aligned}$$

These five sums are the basic quantities you need, and this completes the first step of the calculation.

The next step is to calculate the sum of squared deviations of the x-values and also the sum of products of the deviations of the x- and y-values. You have previously calculated these quantities when calculating a regression line. In addition, the sum of squared deviations of the y-values must also be calculated.

That is, the second step is to calculate

$$\sum (x - \overline{x})^2, \quad \sum (y - \overline{y})^2 \quad \text{and} \quad \sum (x - \overline{x})(y - \overline{y}).$$

Example 4 Calculating a correlation coefficient – step 2

There are eight observations, so $n = 8$.

$$\begin{aligned}\sum(x-\overline{x})^2 &= \sum x^2 - \frac{(\sum x)^2}{n}\\ &= 46\,941.4 - \frac{612.6^2}{8}\\ &= 46\,941.4 - 46\,909.845 = 31.555.\end{aligned}$$

The values calculated above are not rounded as they are going to be used in the next step of the calculation.

$$\begin{aligned}\sum(y-\overline{y})^2 &= \sum y^2 - \frac{(\sum y)^2}{n}\\ &= 24\,039.65 - \frac{437.9^2}{8}\\ &= 24\,039.65 - 23\,969.551\,25 = 70.098\,75.\end{aligned}$$

$$\begin{aligned}\sum(x-\overline{x})(y-\overline{y}) &= \sum xy - \frac{(\sum x)(\sum y)}{n}\\ &= 33\,562.25 - \frac{612.6 \times 437.9}{8}\\ &= 33\,562.25 - 33\,532.1925 = 30.0575.\end{aligned}$$

Step 3 uses the results of step 2 to obtain the correlation coefficient.

Example 5 Calculating a correlation coefficient – step 3

The formula for the correlation coefficient is

$$\text{correlation} = \frac{\sum(x-\overline{x})(y-\overline{y})}{\sqrt{\sum(x-\overline{x})^2 \times \sum(y-\overline{y})^2}},$$

so in this example,

$$\begin{aligned}\text{correlation} &= \frac{30.0575}{\sqrt{31.555 \times 70.098\,75}}\\ &\simeq \frac{30.0575}{47.0315} \simeq 0.639\,09.\end{aligned}$$

The correlation coefficient is therefore 0.64 (to two decimal places).

The procedure for calculating the correlation coefficient is summarised below.

Calculating the correlation coefficient

Given a batch of n linked data pairs, (x, y), the correlation coefficient (r) is obtained as follows:

1. Calculate $\sum x$, $\sum y$, $\sum x^2$, $\sum y^2$ and $\sum xy$.
2. Calculate

$$\begin{aligned}\sum(x-\overline{x})^2 &= \sum x^2 - \frac{1}{n}\left(\sum x\right)^2,\\ \sum(y-\overline{y})^2 &= \sum y^2 - \frac{1}{n}\left(\sum y\right)^2,\\ \sum(x-\overline{x})(y-\overline{y}) &= \sum xy - \frac{1}{n}\left(\sum x\right)\left(\sum y\right).\end{aligned}$$

3. Use the values from step 2 to calculate

$$r = \frac{\sum(x-\overline{x})(y-\overline{y})}{\sqrt{\sum(x-\overline{x})^2 \times \sum(y-\overline{y})^2}}.$$

The next activities provide practice in calculating a correlation coefficient.

Activity 9 Calculating a correlation coefficient

In Examples 3 to 5 a correlation coefficient was calculated for data on student performance in constituencies of one Welsh region, Mid and West Wales. Data for constituencies in other Welsh regions are also available. In particular, data relating to seven constituencies in South Wales West are given in Table 3 below.

Table 3 Results for Key Stage 2 in 2006 and Key Stage 4 in 2011

Percentage achieving benchmark at Key Stage 2 in 2006	Percentage achieving benchmark at Key Stage 4 in 2011
67.2	45.1
76.0	48.3
80.6	65.5
72.3	56.4
71.0	41.9
66.9	40.1
74.0	58.3

(a) Calculate the correlation coefficient between attainment at Key Stage 2 and attainment at Key Stage 4, based on these data.

(b) A scatterplot of the data is given in Figure 5. In the light of this scatterplot, does your answer to part (a) make sense?

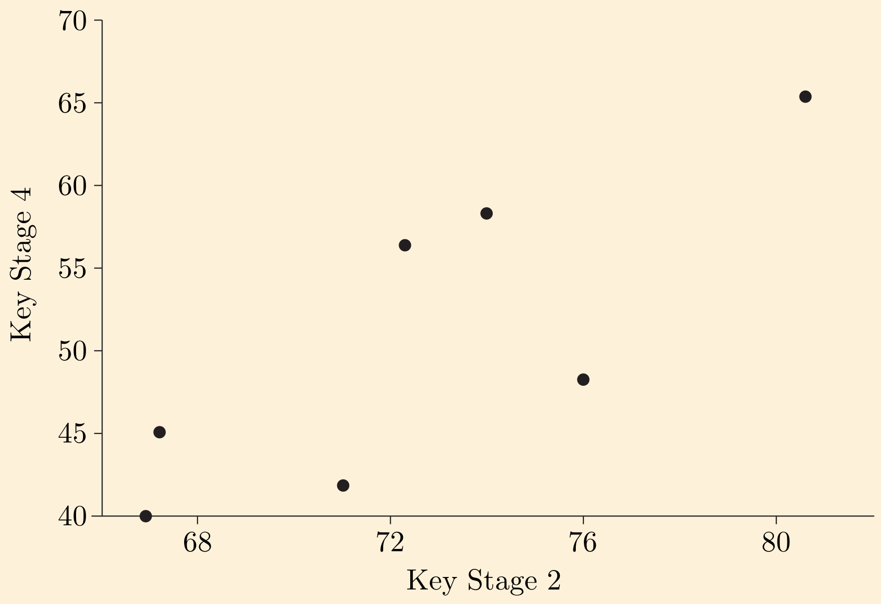
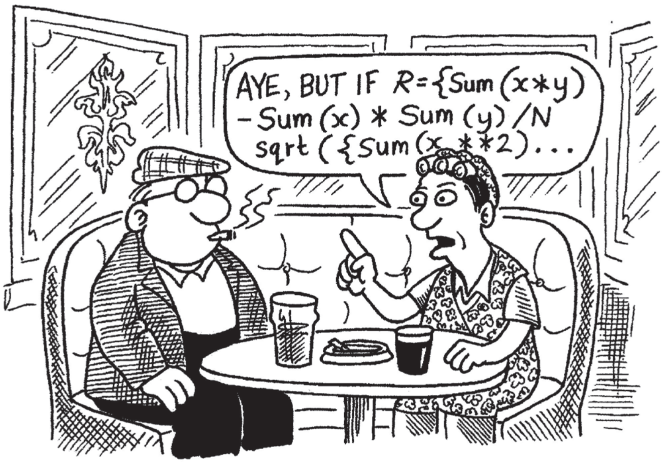


Figure 5 Percentage of children in the assembly constituencies of South Wales West achieving particular benchmarks at Key Stage 2 in 2006 and Key Stage 4 in 2011



'Correlation Street'

Activity 10 Barbecue weather?

Table 4 gives the temperature ($x\,^\circ$ C) on the day before a public holiday in May for the last eight years and the number of barbecue sets a store sold that day (y).

Table 4 Temperature ($x\,^\circ$ C) and sales of barbecue sets (y)

x	16	13	21	17	15	20	15	18
y	4	2	7	4	5	8	4	6

For these data:

$$\sum x = 135, \quad \sum y = 40,$$
$$\sum x^2 = 2329, \quad \sum y^2 = 226, \quad \sum xy = 708.$$

Calculate the correlation coefficient between temperature and number of barbecue sets sold, based on these data.

You have now covered the material related to Screencast 2 for Unit 9 (see the M140 website).

2.3 The sign of the correlation coefficient

We have said that the correlation coefficient is positive when there is a positive linear relationship between x and y, and negative when there is a negative linear relationship between them. To consider why, suppose we have a sample of linked x- and y-values. In some (x, y) pairs, both x and y will be greater than average, sometimes both will be smaller than average, and sometimes one will be larger and the other will be smaller. If x and y are both greater than average, then $(x - \overline{x})$ and $(y - \overline{y})$ are both positive, so

$$(x - \overline{x})(y - \overline{y})$$

is positive. Then that (x, y) pair will contribute a positive amount to the numerator of

$$r = \frac{\sum(x - \overline{x})(y - \overline{y})}{\sqrt{\sum(x - \overline{x})^2 \times \sum(y - \overline{y})^2}}.$$

On the other hand, if x is greater than average while y is smaller than average, then $(x - \overline{x})$ is positive while $(y - \overline{y})$ is negative, so

$$(x - \overline{x})(y - \overline{y})$$

is negative. Hence, that (x, y) pair will contribute a negative amount to the numerator of r. Table 5 shows which (x, y) pairs contribute a positive amount to the numerator of r, and which contribute a negative amount.

Table 5 The sign of contributions to the correlation coefficient

x-value	y-value	$x - \overline{x}$	$y - \overline{y}$	$(x - \overline{x})(y - \overline{y})$
greater than $\overline{x}$	greater than $\overline{y}$	positive	positive	positive
less than $\overline{x}$	less than $\overline{y}$	negative	negative	positive
greater than $\overline{x}$	less than $\overline{y}$	positive	negative	negative
less than $\overline{x}$	greater than $\overline{y}$	negative	positive	negative

Pictorially, Table 5 corresponds to the following areas on the scatterplot in Figure 6.

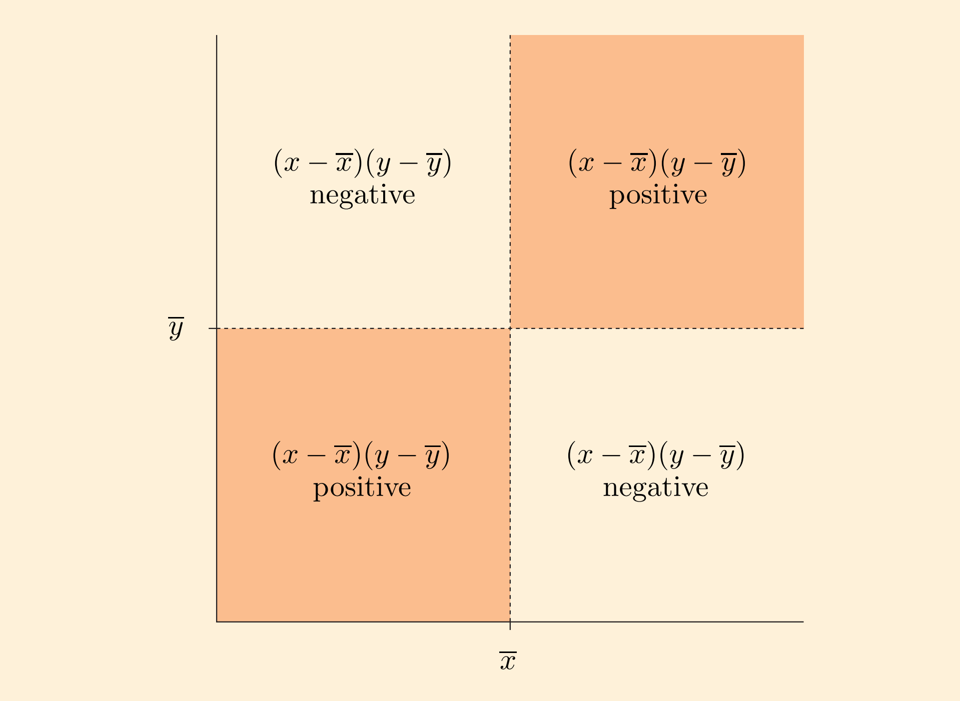

$(x - \overline{x})(y - \overline{y})$
negative
$(x - \overline{x})(y - \overline{y})$
positive
$\overline{y}$
$(x - \overline{x})(y - \overline{y})$
positive
$(x - \overline{x})(y - \overline{y})$
negative
$\overline{x}$

Figure 6 Contributions to the numerator of the correlation coefficient

Now if most data points are in the shaded areas (top right and bottom left), then

- most points will contribute a positive amount to the numerator of r
- the regression line through the data will slope upwards, and there is a positive linear relationship between x and y.

On the other hand, if most data points are in the non-shaded areas (top left and bottom right), then

- most points will contribute a negative amount to the numerator of r
- the regression line through the data will slope downwards, and there is a negative linear relationship between x and y.

Finally, note that because the denominator of the formula for r is always positive, the sign of the correlation is always the same as the sign of the numerator. Hence, the correlation coefficient is positive when there is a positive linear relationship and negative when there is a negative linear relationship.

Example 6 Deducing the sign of the correlation coefficient

In Figure 7 a dataset consisting of 20 observations is plotted. Also, the regions where $(x - \overline{x})(y - \overline{y})$ is positive are shaded.

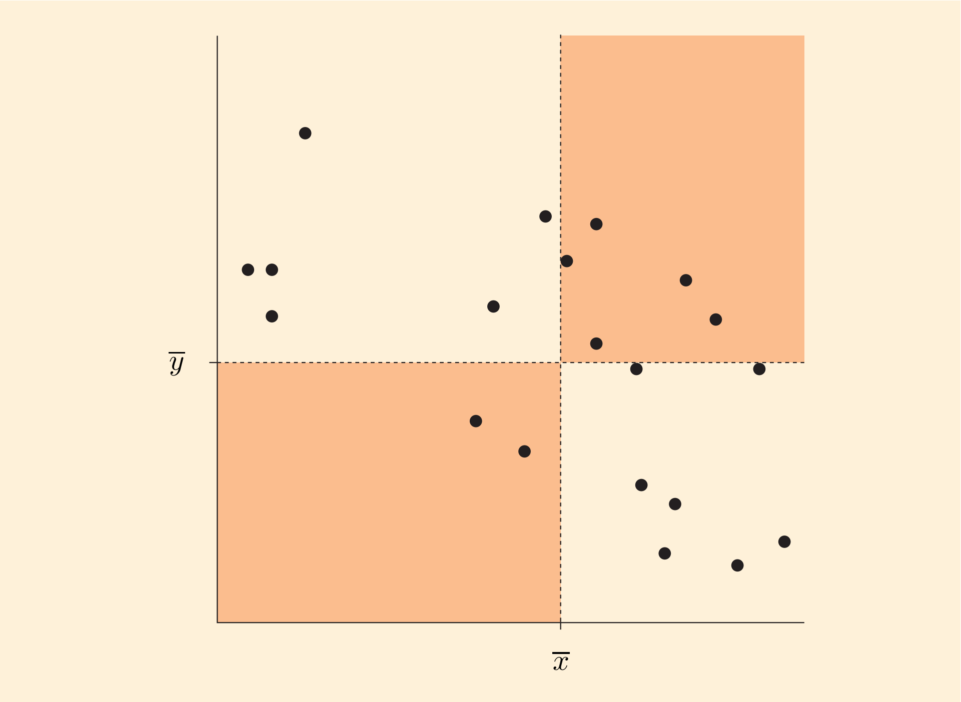


Figure 7 A dataset of 20 observations

Notice that, in this plot, just seven of the twenty points lie in the shaded region. These will give a positive contribution to the numerator of the correlation coefficients. However, they are outweighed by the points giving a negative contribution to the numerator, as there are thirteen such points. (Also, points (x, y) make a greater contribution to $\sum(x - \overline{x})(y - \overline{y})$ as they get further from $(\overline{x}, \overline{y})$, and the points in the non-shaded regions tend to be further from $(\overline{x}, \overline{y})$ than the points in the shaded regions.)

So overall the correlation coefficient will be negative (as it should be for variables that are clearly negatively related). In fact the correlation coefficient turns out to be $r = -0.57$.

Activity 11 Deducing the sign of another correlation coefficient

A different dataset of 20 observations is plotted in Figure 8. The regions where $(x - \overline{x})(y - \overline{y})$ is positive are shaded.

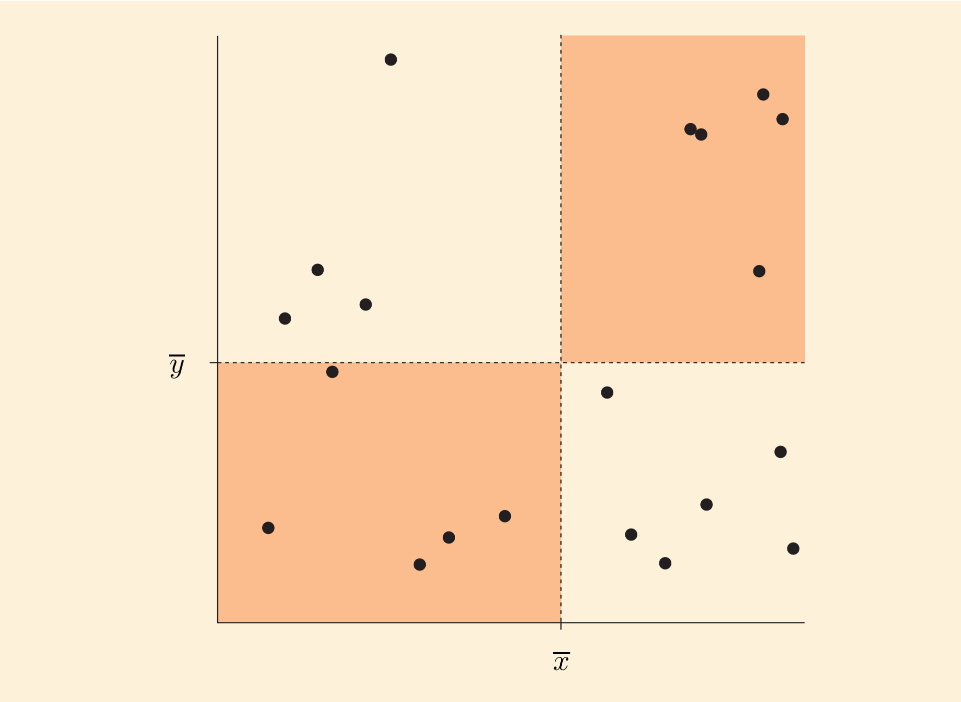


Figure 8 Another dataset of 20 observations

(a) On the basis of this plot, does there appear to be a positive relationship, a negative relationship or no relationship?

(b) By considering contributions to the numerator of the correlation coefficient, is the correlation coefficient likely to be positive, negative or close to zero?

2.4 Computer work: correlation coefficients

As you have seen in Subsection 2.2, calculating the correlation coefficient by hand is straightforward. However, the calculations are tedious, particularly when the dataset is not small. So it is far more common to use a computer instead.

In this subsection you will learn how to calculate a correlation coefficient using Minitab. You should now turn to Chapter 9 of the Computer Book and work through Subsection 9.1.

Exercises on Section 2

Exercise 2 Ordering correlations

Order the following correlations from strongest to weakest.

$$+0.25 \quad +0.80 \quad -0.30 \quad -0.99 \quad +0.04$$

Exercise 3 Estimating more correlations

Figure 9 shows a scatterplot of the best times for the 200-metre sprint and 100-metre sprint by male UK sprinters in 2011.

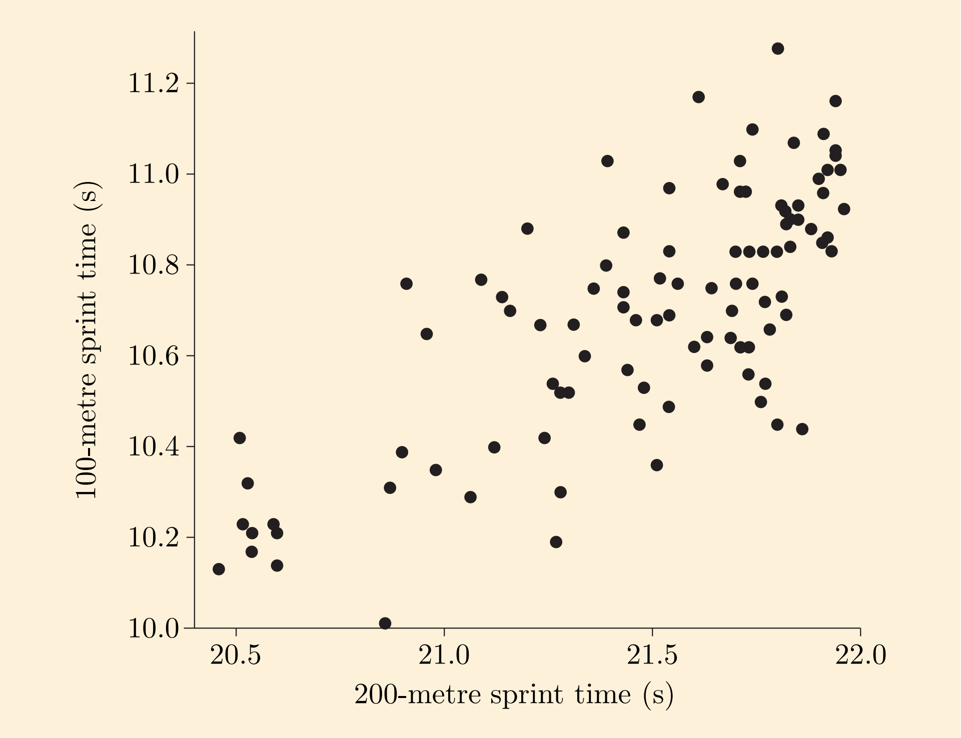


Figure 9 Times for male UK sprinters in 2011

Using this plot, estimate the correlation between the best time on the 200-metre sprint and the best time on the 100-metre sprint.

Exercise 4 Calculating another correlation coefficient

The data for the top nine sprinters in the 200-metre sprint plotted in Figure 9 are given in Table 6 below.

Table 6 Best performances of top nine sprinters

Name	200-metre time (s)	100-metre time (s)
Harry Aikines-Aryeety	20.46	10.13
Leon Baptiste	20.51	10.42
James Ellington	20.52	10.23
Richard Kilty	20.53	10.32
Danny Talbot	20.54	10.21
Christian Malcolm	20.54	10.17
James Alaka	20.59	10.23
Marlon Devonish	20.60	10.14
Luke Fagan	20.60	10.21

Marlon Devonish in action

Using these data, calculate by hand the correlation coefficient between the 200-metre time and the 100-metre time for this group of sprinters.

Compare the correlation coefficient for these nine sprinters with the correlation coefficient in Exercise 3.

Exercise 5 Deducing the sign of yet another correlation coefficient

A dataset of 20 observations is plotted in Figure 10. The regions where $(x-\overline{x})(y-\overline{y})$ is positive are shaded.

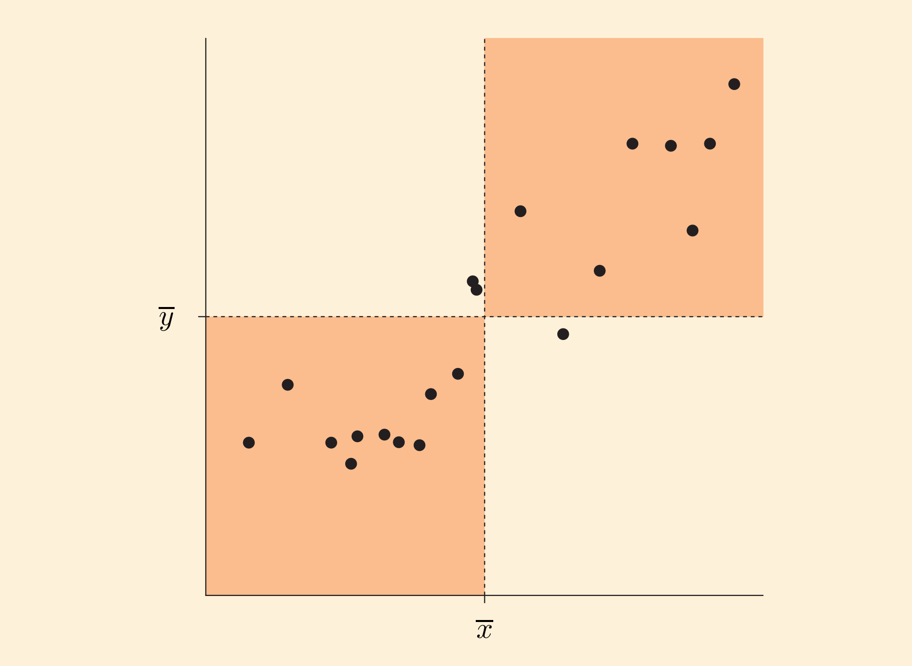


Figure 10 A dataset of 20 observations

(a) On the basis of this plot, does there appear to be a positive relationship, a negative relationship or no relationship?

(b) By considering contributions to the numerator of the correlation coefficient, is the correlation coefficient likely to be positive, negative or close to zero?

3 More on the correlation coefficient

In Section 2 you saw that the correlation coefficient is a measure of the strength of a relationship. There you saw that the correlation coefficient takes a value between $+1$ and -1, with $+1$ indicating an exact positive relationship, -1 indicating an exact negative relationship, and 0 indicating no relationship. In this section we shall look further at what aspects of the data affect the value of the correlation coefficient.

3.1 Specifying the variables

In Figures 2.1 to 2.1 in Subsection 2.1, which consisted of made-up data, you may have noticed that no scales were included. In general, this is very bad practice for a graph, but in this case there was a specific reason.

> The correlation coefficient does not depend on the scales of the axes. It only reflects the pattern of the points.

Let us see what this means by looking at the following example.

Example 7 Pass rates in English and Mathematics

Figure 11 shows data on pass rates in English and Mathematics qualifications at Key Stage 4 for the 100 English secondary schools in our sample. You can see that there is a definite positive correlation between the pass rates in English and Mathematics in the schools; the correlation coefficient turns out to be $+0.70$.

Now, suppose that instead of plotting the pass rates in English as given, we plot them as rates (percentages) over 50%. So a pass rate of 80% in English will appear as $+30$ and a pass rate of 40% will appear as -10. Also, we could plot the pass rates in Mathematics as proportions more than 0.5. So a pass rate of 80% in Mathematics will appear as $+0.3$ and a pass rate of 40% will appear as -0.1. Then the scatterplot appears as in Figure 12 and though the numbers on the axes are different, the pattern of points is exactly the same as in Figure 11. The correlation coefficient does not change; it is still $r = +0.70$.

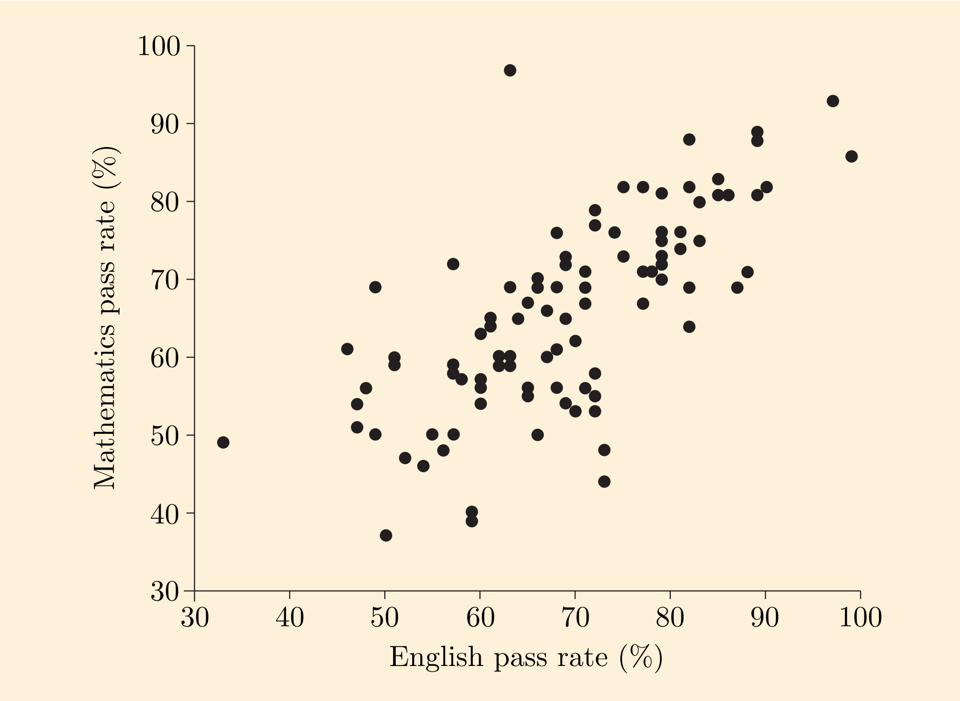


Figure 11 Pass rates in English and Mathematics at Key Stage 4 for 100 non-selective schools

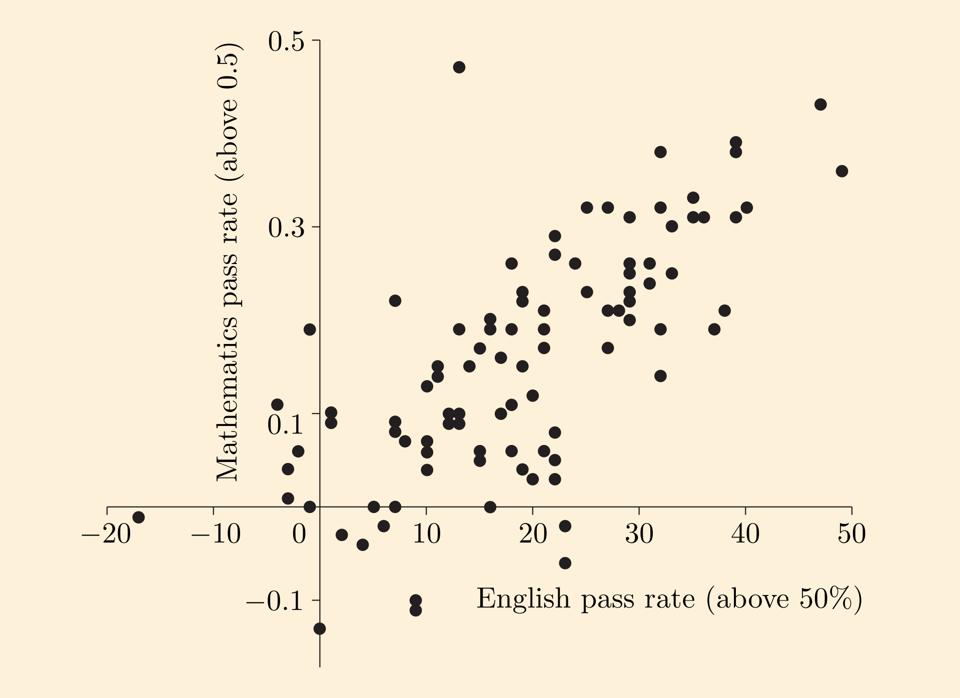


Figure 12 Pass rates in English and Mathematics at Key Stage 4 for 100 non-selective schools

So the coefficient is not changed either by changing the position of one or both axes or by recording measurements in different units.

In Figures 11 and 12, the pattern of points looks exactly the same, although the numbers are different. However, it also does not make any difference if one of the scales is changed so that the pattern looks different. This has been done in Figure 13. The correlation coefficient is still $+0.70$ in this scatterplot. It is the amount of scatter about a straight line that is important, not how steep the line is. In all these figures, the points could all be included inside a fairly narrow region.

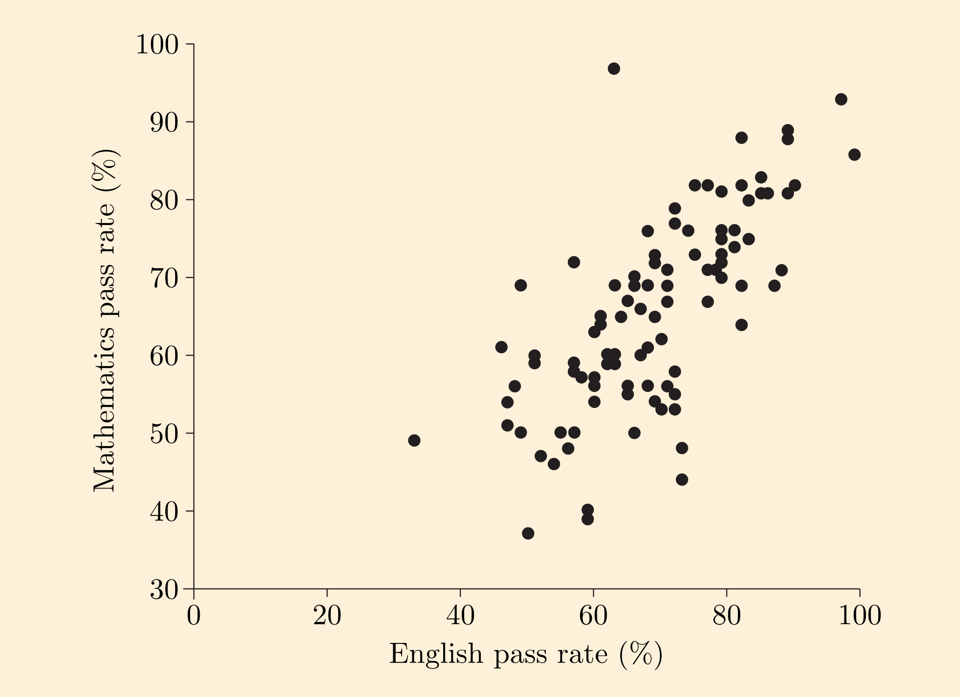


Figure 13 Pass rates in English and Mathematics at Key Stage 4 for 100 non-selective schools

Another factor which does not affect the value of the correlation coefficient is which variable is plotted on the x-axis. In this example, there is no obvious reason why either one of the variables is dependent on the other, and in Figure 14, the pass rates for Mathematics are plotted on the x-axis, and the pass rates for English on the y-axis. Although the pattern of points in the scatterplot looks different, the value of the correlation coefficient is still unchanged at $+0.70$.

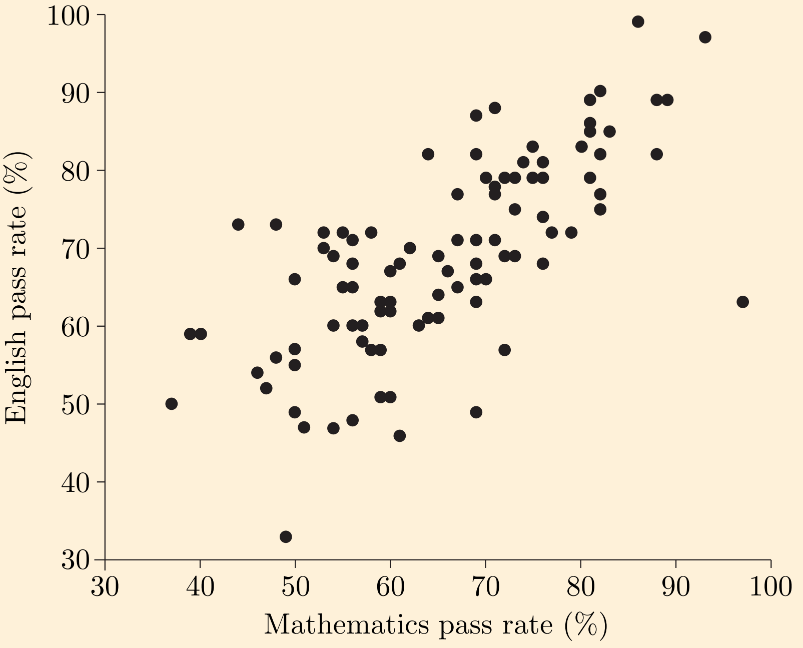


Figure 14 Pass rates in English and Mathematics at Key Stage 4 for 100 non-selective schools

Thus we have the following.

The correlation coefficient does not depend on which variable is plotted on the vertical axis and which is plotted on the horizontal axis.

Activity 12 Same correlation?

Suppose that the correlation is $+0.56$ for the following pair of variables: a car's weight (in tonnes) and the miles per gallon it can achieve. Which of the following pairs of variables would have exactly the same correlation as this pair?

(a) A car's weight (in kilograms) and the miles per gallon it can achieve.

(b) A car's length (in metres) and the miles per gallon it can achieve.

(c) A car's weight (in tonnes) and the kilometres per litre it can achieve.

(d) A car's weight (in tonnes) and its fuel consumption (measured in litres per kilometre).

(e) The miles per gallon over 50 miles per gallon that a car can achieve, and a car's weight (in tonnes over 1 tonne).

3.2 The shape of the relationship

In Subsection 2.1 you learned that if there is no relationship between two variables, the correlation coefficient is equal to, or very close to, zero. Does that mean that when the correlation coefficient between variables is close to zero, there is no relationship between the variables? The answer is: *not necessarily*.

Example 8 A curve with zero correlation

Figure 15 is a scatterplot of some made-up data. For these data it also turns out that $r = 0.0$.

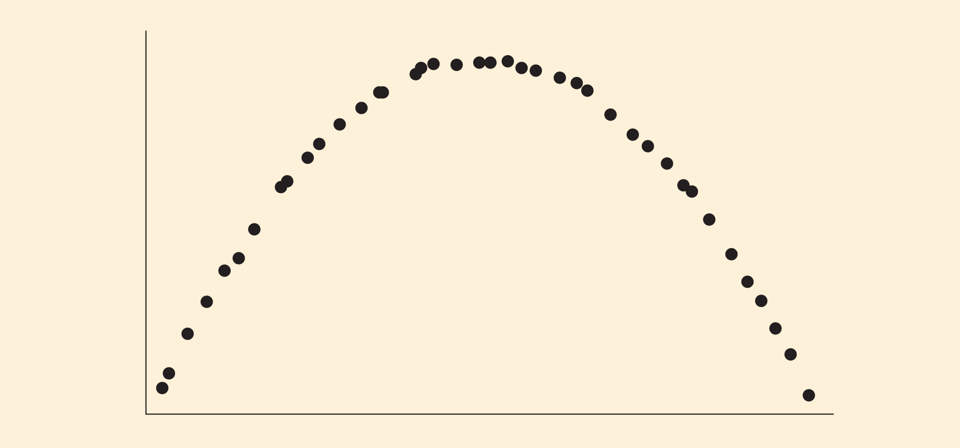


Figure 15 A scatterplot of some data

However, notice that despite this zero correlation, there is clearly a very strong relationship between the two variables.

So why does the strong relationship between the two variables plotted in Example 8 have a correlation coefficient of zero? The answer is that the correlation coefficient r only measures the extent of the *linear* relationship between two variables. It does not take account of a non-linear pattern in the relationship.

A correlation coefficient close to zero does *not* imply that there is no relationship, merely that there is not a linear relationship.

However, not all non-linear relationships will have a correlation coefficient close to zero, as the next example shows.

Example 9 Key Stage 2 pass rates

In Figure 16, the percentage of pupils in England passing their Key Stage 2 test in English at Level 4 or above between 1995 and 2011 is plotted. Notice from the plot that there seems to be a close positive relationship between percentage pass rate and year: a smooth curve could be drawn on the graph that is very close to all the points. However, the relationship is not linear. In the late 1990s, the pass rate went up rapidly. In contrast, the pass rate only went up by a few percentage points in the period 2004 to 2011.

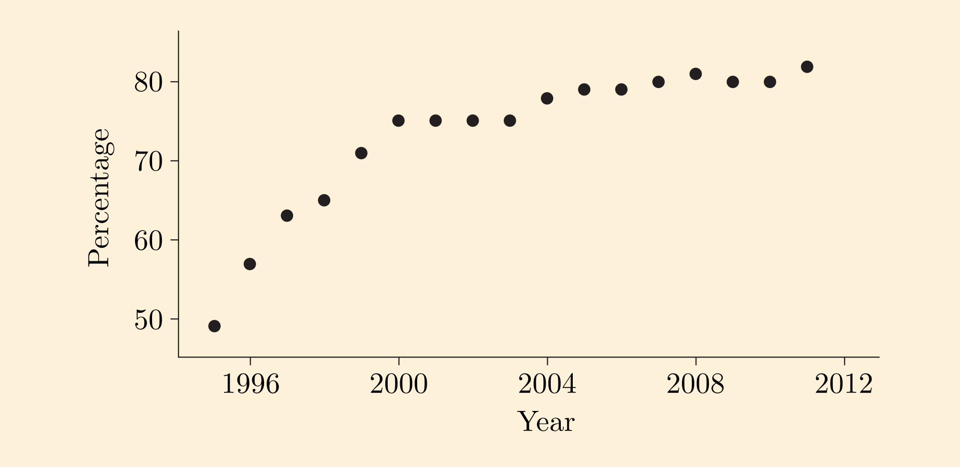


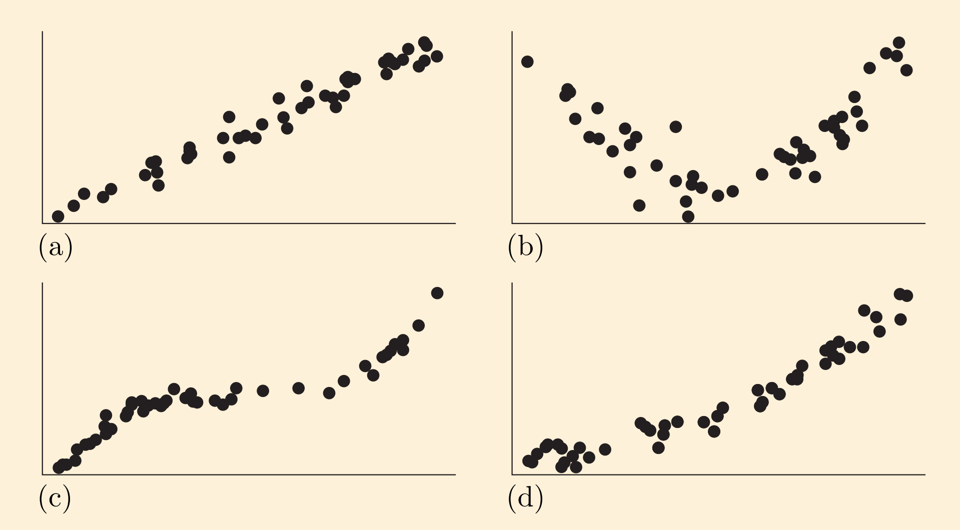
Figure 16 Percentage of Key Stage 2 students at Level 4 or above in English over time

For these data, $r = +0.88$. This is quite close to $+1$. However, such a close relationship between two quantities would give a much bigger value of r – between $+0.95$ and $+0.99$ – if the relationship were linear.

The correlation coefficient just measures the strength of a linear relationship. When a non-linear relationship is generally positive, the correlation coefficient will be positive but not as large as for a similarly strong positive *linear* relationship. Equally, when a non-linear relationship is generally negative, the correlation coefficient will be negative, but not as negative as for a similarly strong negative linear relationship.

Activity 13 Representative correlation?

For each of the scatterplots shown below, state whether the correlation coefficient is likely to be a good indication of the strength of the relationship between the two variables.



3.3 Outliers and influential points

The purpose of this subsection is to show how one or two data points can sometimes have an overriding effect on a correlation coefficient. It can happen in two possible ways. First, look at Figure 17, which gives a scatterplot of the number of boys and the number of girls reaching the end of Key Stage 4 in our sample of 100 schools.

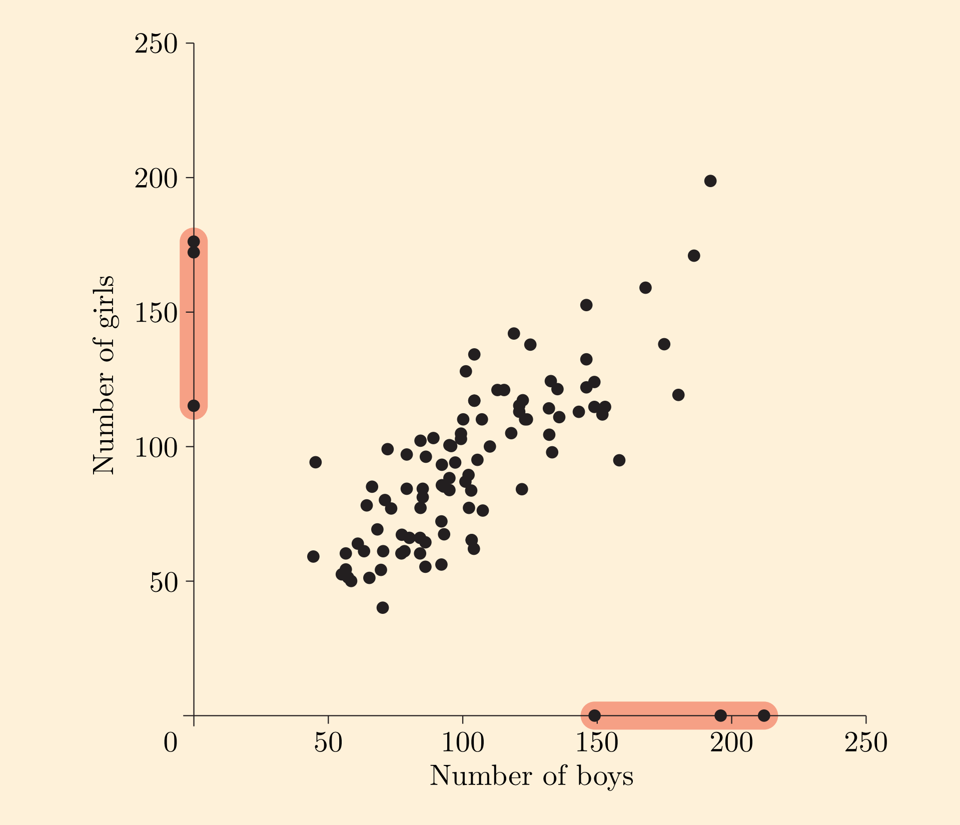


Figure 17 Numbers of boys and girls ending Key Stage 4

In this scatterplot two groups of points have been highlighted: a group of three schools where there were no girls and a group of three schools where there were no boys. These schools do not follow the pattern of the other schools, which tended to have similar numbers of boys and girls. These points are therefore outliers. In fact, these six schools are single-sex schools, whereas all the other schools are mixed-sex schools. The point to notice here is that when the seven single sex schools are included, the correlation coefficient is $+0.23$; when they are omitted, the coefficient is $+0.74$. So, if only mixed-sex schools are included, the relationship between the numbers of boys and girls in schools ending Key Stage 4 is strong and positive; but if all schools are included, the relationship is weak.

The second way in which a very small number of points can exert an overriding effect is shown in Figure 18. The x-variable is the percentage of children who achieved the Key Stage 2 benchmark in 2006 and the y-variable is the percentage of children who achieved the Key Stage 4 benchmark in 2011, both introduced in Example 3 (Subsection 2.2).

The ringed point refers to one particular constituency, Llanelli. You can see that the ringed point is **remote** from the rest; the percentage of children achieving the Key Stage 2 benchmark and the percentage achieving the Key Stage 4 benchmark were both noticeably lower in Llanelli than in the other constituencies.

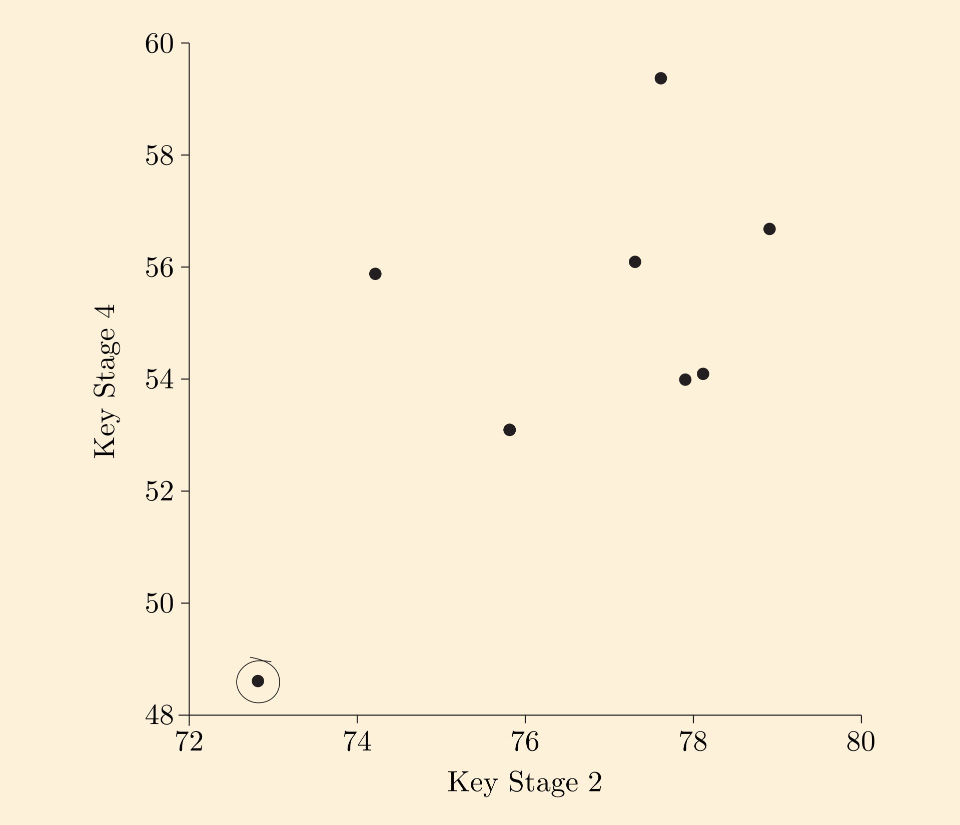


Figure 18 Percentage of children achieving particular benchmarks at Key Stage 2 in 2006 and Key Stage 4 in 2011

The point cannot be considered an outlier because it does not appear to be out of line with the remaining points. However, without that point, the correlation would be much less pronounced; the coefficient is reduced from $+0.64$ to $+0.18$ if the point for Llanelli is omitted. Such a point is sometimes called an **influential point**. An influential point has an x-value and/or a y-value that is a long way from those of the other points.

It is worth noting that removing an influential point *reduces* the amount of correlation (i.e. moves the correlation coefficient closer to zero), whereas removing an outlier usually *increases* it (moves the coefficient towards $+1$ or -1).

Activity 14 Spotting outliers and influential points

Data are also available for other regions of Wales, and data for the eight constituencies in South Wales East are plotted in Figure 19.

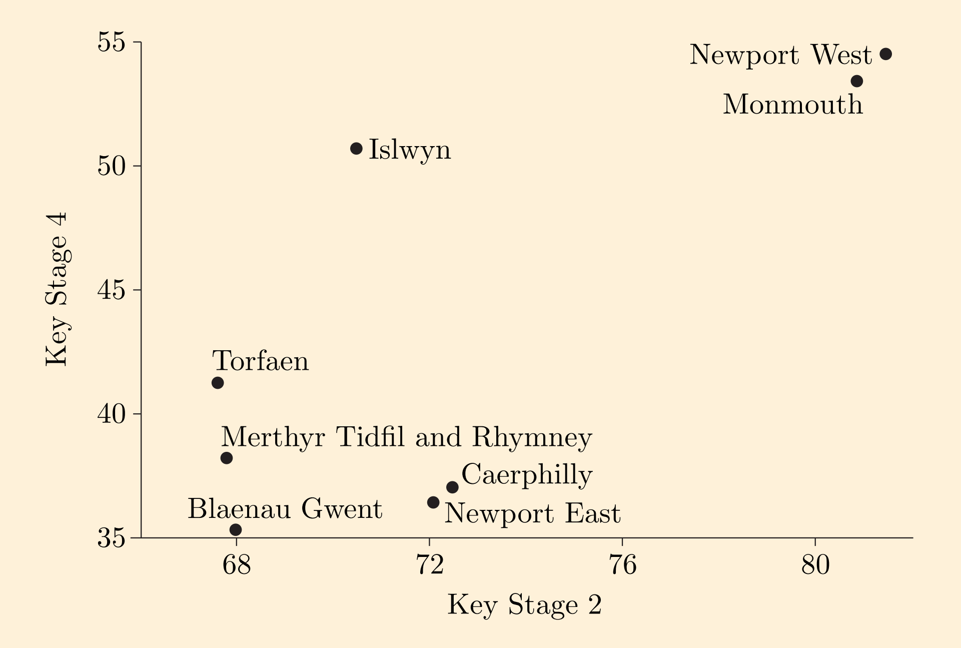


Figure 19 Percentage of children in the assembly constituencies of South Wales East achieving particular benchmarks at Key Stage 2 in 2006 and Key Stage 4 in 2011

Identify any points which are outliers or influential points, and make rough guesses of the correlation coefficients both including and excluding these points.

You have now covered the material related to Screencast 3 for Unit 9 (see the M140 website).

3.4 Correlation and causation

The final point we shall note about correlations is the following.

Correlation is *not* causation.

That is, just because two variables are correlated does not mean that one causes the other.

Example 10 Pass rates in English and Mathematics

In Example 7 (Subsection 3.1) you saw that there is a strong positive correlation between a school's pass rate in English and its pass rate in Mathematics. However, does this mean that a high rate pass rate in Mathematics causes a high pass rate in English?

That is, can the relationship be summarised by the diagram in Figure 20? (In such diagrams the arrow denotes a **causal relationship**.)

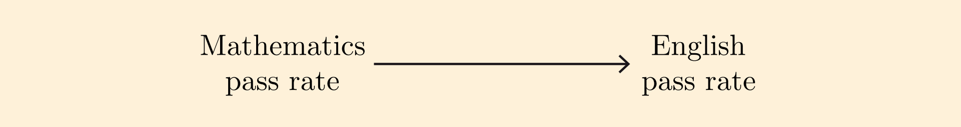


Figure 20 A possible relationship between pass rates in English and Mathematics

No, it does not. It could be the other way round – a high pass rate in English might cause a high pass rate in Mathematics, as shown in Figure 21. The correlation would be the same.

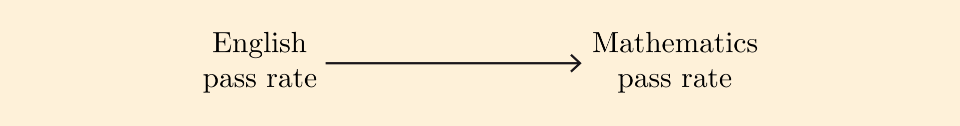


Figure 21 Another possible relationship between pass rates in English and Mathematics

However, it is more likely that the actual causal relationships are closer to one of those shown in Figures 22 and 23. That is, the relationship between the pass rates in English and Mathematics are due to some other causal relationship, such as the ability of students and/or the quality of the school.

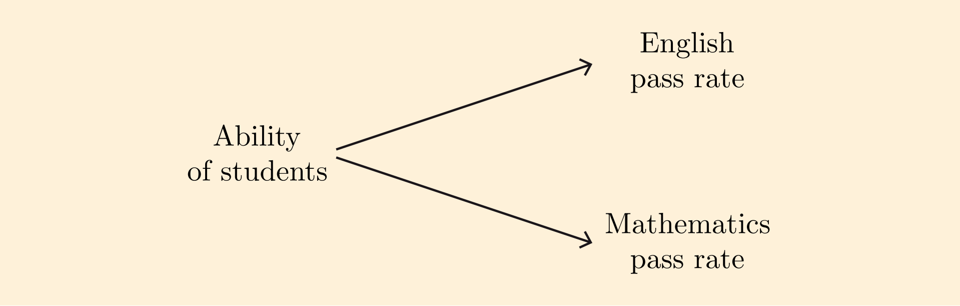


Figure 22 Another possible relationship between pass rates in English and Mathematics

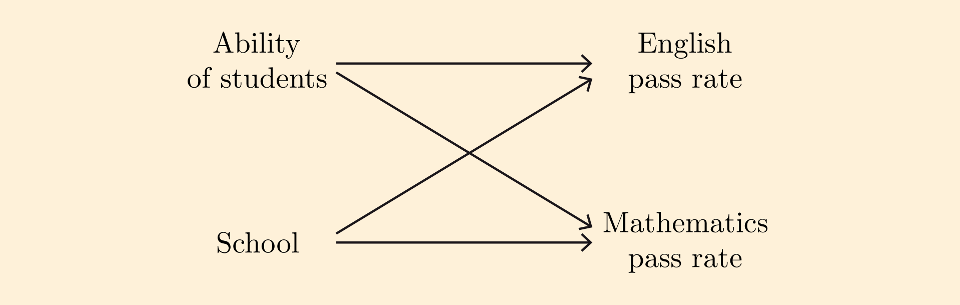
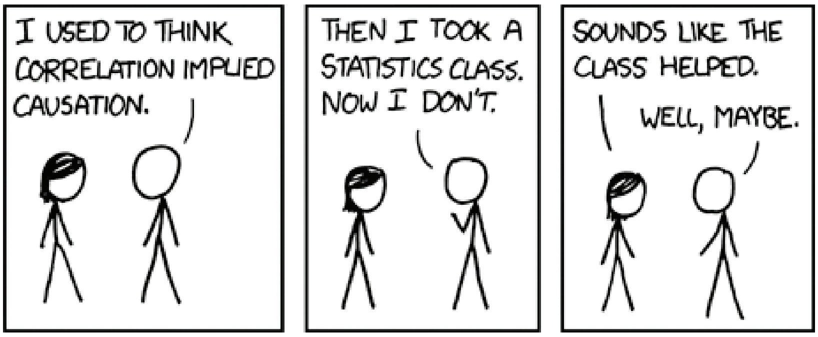


Figure 23 Yet another possible relationship between pass rates in English and Mathematics

From just the correlation coefficient, it is impossible to tell which of these situations represents the true causal relationships – if any of them do! They can all lead to a high positive correlation between the English and Mathematics pass rates.


I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.

Activity 15 Does chocolate make you clever?

In November 2012, the BBC News website reported on a study about a country's chocolate consumption and the country's number of Nobel prize laureates. A plot from that report is reproduced in Figure 24.

The correlation coefficient is shown on Figure 24; its value is $+0.791$. Does this correlation coefficient provide evidence that high chocolate consumption makes a country's inhabitants clever?

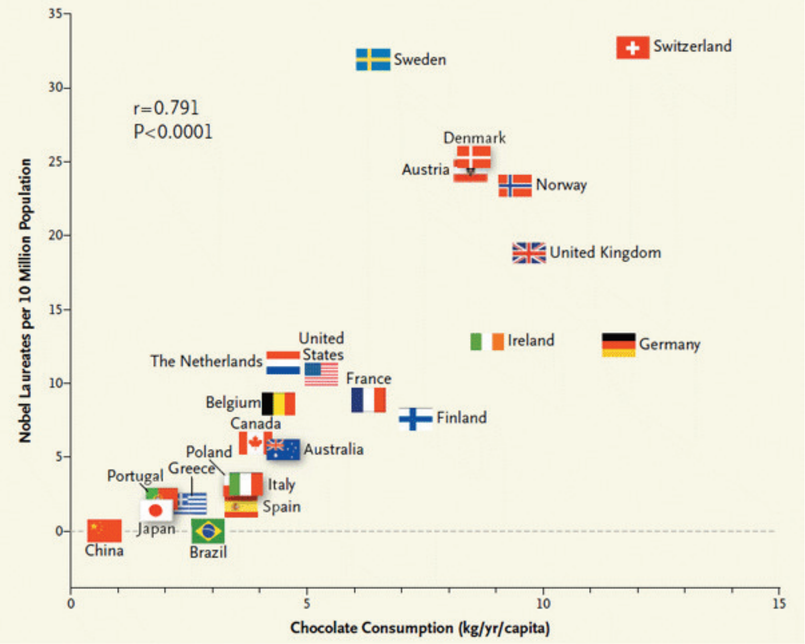


Figure 24 Chocolate consumption and numbers of Nobel prize laureates

Exercises on Section 3

Exercise 6 Does the correlation coefficient change?

In Activity 10 (Subsection 2.2) you calculated the correlation coefficient between the temperature (x) on the day before a public holiday and the number of barbecue sets a store sold that day (y). This correlation coefficient turned out to be $+0.91$.

(a) If the number of barbecue sets that were sold had been recorded as x and temperature had been recorded as y, would the correlation coefficient be different?

(b) The temperature was recorded in *degrees Celsius* (° C). Another temperature scale is the *Kelvin scale* (K). A temperature measured in degrees Celsius can be converted to kelvins simply by adding 273.15. For example, a temperature of $16\,^\circ$ C is $289.15\,$K. If temperatures had been measured on the Kelvin scale, rather than in degrees Celsius, would the correlation coefficient be different?

(c) Another scale that is used to record temperature is the *Fahrenheit scale* (° F). To convert a temperature from Celsius to Fahrenheit, you first multiply the temperature by 1.8, and then you add 32 to the result. For $16\,^\circ$ C the calculation is as follows: $(16 \times 1.8) + 32 = 60.8$. So $16\,^\circ$ C is the same as

60.8 ° F. If temperatures had been measured in degrees Fahrenheit, rather than in degrees Celsius, would the correlation coefficient be different?

Exercise 7 Estimating yet more correlations

In Exercise 4 (Section 2) you were asked to calculate the correlation coefficient between the best time for the 200-metre sprint and the best time for the 100-metre sprint, based on nine UK sprinters. A plot of these times is given in Figure 25.

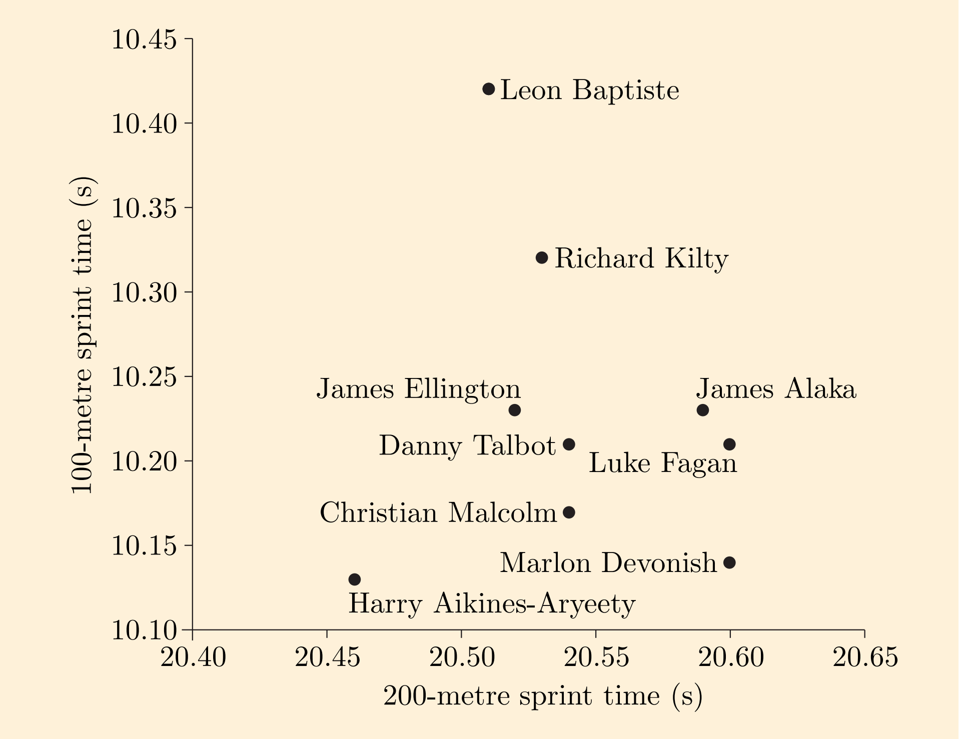


Figure 25 Times for top UK male 200-metre sprinters in 2011

Identify any points which are outliers or influential points and make rough guesses of the correlation coefficients both including and excluding these points.

Exercise 8 Investigating the tastiness of fruit

An assessment of the tastiness of fruit, along with how easy they are to eat, is plotted on Figure 26 below. Assuming that these assessments provide a realistic comparison of fruit, answer the following questions.

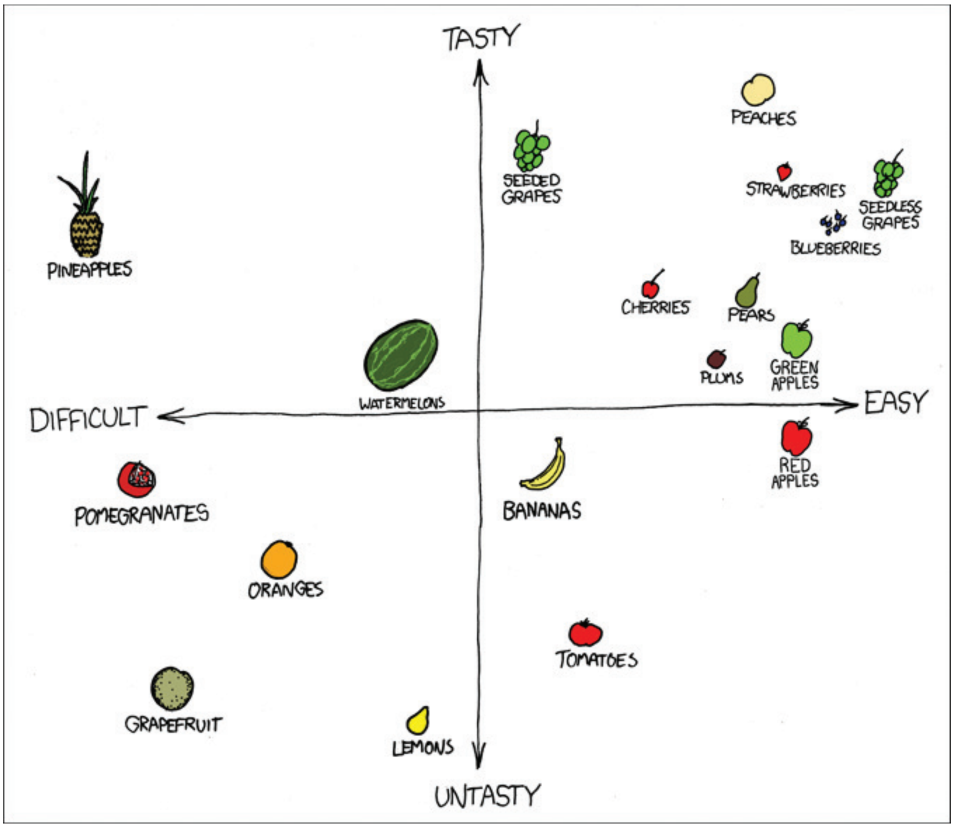


Figure 26 Taste versus difficulty in eating

(a) Do any fruits in Figure 26 appear to be outliers? Do any appear to be influential points? Justify your answers.

(b) After looking at Figure 26, one person said: ‘So this means being easy to eat makes a fruit more tasty’. Is this person’s conclusion appropriate?

4 Confidence intervals from z-tests

An important use of statistical methods is to make estimates. In our everyday lives we commonly give estimates in the form of intervals, such as ‘the work will take me 30 to 40 minutes to do’ or ‘it will cost \$50 to \$60’. How such intervals should be interpreted is not always completely clear – is it certain that the work will take somewhere between 30 and 40 minutes, or simply very likely that it will be completed within that time? We next consider ways of giving estimates as intervals that have precise interpretations. In this section we obtain interval estimates that are related to z-tests – these are readily calculated without a computer. In Section 5 we consider interval estimates in the context of least squares regression, when a computer is generally used to calculate the intervals.

4.1 Estimating the population mean

In Section 5 of Unit 7 you were introduced to the one-sample z-test. The test is used to examine whether a population mean has a specified value, say A. It involves setting up the hypothesis that the population mean is equal to A, and then using sample data to decide whether or not to reject the hypothesis at the 5% significance level. If we do *not* reject the hypothesis, we conclude that the sample could well have come from a population with mean A; whereas if we *do* reject the hypothesis, we conclude that, based on the sample, the population mean is probably not equal to A.

In neither case do we say much about what the true value of A is. The sample mean is an estimate of A, but often it would be helpful to have an interval that gives the range of values that A might plausibly have. We shall obtain a method of constructing such an interval. The method is derived from the one-sample z-test.

In Section 1 of this unit we had a random sample of 100 English state secondary schools. The value of P_{KS4} (the GCSE headline figure) was determined for each school. These values have a mean of 50.3 and a standard deviation of 17.19.

Example 11 z-test of whether P_{KS4} equals 45.0

Suppose we want to test, at the 5% significance level, whether the mean value of P_{KS4} in English state schools equals 45.0%.

Then we let μ denote the mean of P_{KS4} in English state schools and set up the hypotheses

$$H_0: \mu = 45.0 \quad \text{and} \quad H_1: \mu \neq 45.0.$$

The population standard deviation is unknown and the size is large (over 25), so the test statistic is

$$z = \frac{\overline{x} - A}{\text{ESE}}, \quad \text{where ESE} = \frac{s}{\sqrt{n}}.$$

Here, $A = 45.0$, $\overline{x} = 50.3$, $n = 100$ and $s = 17.19$. Hence,

$$z = \frac{50.3 - 45.0}{17.19/\sqrt{100}} \simeq 3.08.$$

The critical values for a z-test at the 5% significance level are -1.96 and 1.96. As 3.08 is greater than 1.96, the null hypothesis is rejected at the 5% significance level. Thus there is moderate evidence that the mean value of P_{KS4} is not 45.0%.

The hypothesis test enables us to conclude that the population mean is probably not equal to 45.0%, but it does not actually give an estimate for the mean. If we want to estimate the population mean by a single value, then the sample mean is the obvious choice, and it can give a good idea of what the population mean is likely to be. Estimates such as this, which consist of a single number, are called **point estimates**.

They are not entirely satisfactory as estimates because, without other information, we have no indication of their accuracy. A more satisfactory approach is to specify a range of likely values for the population mean, by finding two numbers between which the mean is likely to lie. We shall use the sample values to find these two numbers. Such a range of numbers is called an **interval** and is said to provide an **interval estimate** of the population mean.

In Subsection 4.2 we shall explore the most widely used form of interval estimate. It is called a **confidence interval** and it is based on many of the ideas that were introduced for hypothesis testing.

4.2 An interval estimate for the mean

We are now going to consider a whole series of hypothesis tests of the form

$$H_0: \mu = A \quad \text{and} \quad H_1: \mu \neq A.$$

In Example 11 (Subsection 4.1) we rejected the hypothesis that the population mean is equal to A when we put $A = 45\%$. In Activity 16 you are asked to repeat the test for other values of A.

Activity 16 z-test of whether P_{KS4} equals 47.0, 53.0 and 55.0

The mean and standard deviation of P_{KS4} in a sample of 100 schools were equal to 50.3 and 17.19, respectively. Carry out separate one-sample z-tests at the 5% significance level for each of the following hypotheses.

(a) The population mean is 47.0%.

(b) The population mean is 53.0%.

(c) The population mean is 55.0%.

In Activity 16 we tested the hypothesis $H_0: \ \mu = A$ for $A = 47.0$, $A = 53.0$ and $A = 55.0$. Following on from the activity, we might ask the question: 'What values of A would be rejected in a test at the 5% significance level, and what values would we fail to reject?'

The question is important because we want an interval that contains the plausible values of A. For this, we might form the interval that consists of *those values of A that are not rejected at the 5% significance level.* This interval is called the **95% confidence interval.** A 99% confidence interval is defined in a similar way, but using a 1% significance level in the hypothesis tests.

Confidence intervals

A 95% confidence interval for μ includes all values of A for which we cannot reject H_0 at the 5% significance level.

A 99% confidence interval for μ includes all values of A for which we cannot reject H_0 at the 1% significance level.

Performing lots of hypothesis tests seems a laborious way of determining a confidence interval, so we ask: *is there a straightforward way of finding confidence intervals for a population mean?*

Well, in Example 11 and Activity 16 the null hypothesis was rejected if (i) z was greater than or equal to 1.96, or (ii) z was less than or equal to -1.96. The value of z was obtained by calculating

$$z = \frac{50.3 - A}{17.19/\sqrt{100}} = \frac{50.3 - A}{1.719}.$$

So when is the null hypothesis not rejected? It is not rejected by (i) if z is less than 1.96, so when

$$\frac{50.3 - A}{1.719} < 1.96,$$

which implies

$$50.3 - A < 1.96 \times 1.719,$$

which in turn implies

$$50.3 - 1.96 \times 1.719 < A.$$

Similarly, it is not rejected by (ii) if z is greater than -1.96, so when

$$-1.96 < \frac{50.3 - A}{1.719},$$

which implies

$$-1.96 \times 1.719 < 50.3 - A,$$

giving

$$A < 50.3 + 1.96 \times 1.719.$$

Thus the value of A is rejected by neither (i) nor (ii) when

$$50.3 - 1.96 \times 1.719 < A < 50.3 + 1.96 \times 1.719,$$

that is, when (rounding to one decimal place)

$$46.9 < A < 53.7.$$

This gives the range of values for A for which we do not reject the null hypothesis at the 5% significance level. Thus it is also our 95% confidence interval for the population mean. The interval consisting of all numbers from 46.9 to 53.7 is written

$$(46.9, 53.7).$$

So we would write that $(46.9\%, 53.7\%)$ is the 95% confidence interval for the mean P_{KS4} in English state schools.

If a hypothesis is rejected at the 5% significance level, it means that the test statistic was one of the 5% most extreme values. Thus, if the null hypothesis is true, there is a probability of $1 - 0.05 = 0.95$ or 95% that the sample we select will *not* give a test statistic in one of the extreme tails of the distribution. This is the area we consider for our confidence interval, so this is why it is called a 95% confidence interval.

We could start with

$$z = \frac{\overline{x} - A}{s/\sqrt{n}} \quad \text{rather than} \quad z = \frac{50.3 - A}{17.19/\sqrt{100}}$$

and perform the same operations. We would find that a value of A is rejected by neither (i) nor (ii) at the 5% significance level when

$$\overline{x} - 1.96 \times s/\sqrt{n} < A < \overline{x} + 1.96 \times s/\sqrt{n}.$$

This gives a procedure for calculating the 95% confidence interval for a population mean. Replacing 1.96 by 2.58 gives the 99% confidence interval.

Calculating 95% and 99% confidence intervals for a population mean

Suppose the sample size is n, the sample mean is $\overline{x}$, and the sample standard deviation is s.

- Calculate the estimated standard error: ESE $= s/\sqrt{n}$.
- The 95% confidence interval for the population mean is

  $$(\overline{x} - 1.96\,\text{ESE},\ \overline{x} + 1.96\,\text{ESE}).$$

- The 99% confidence interval for the population mean is

  $$(\overline{x} - 2.58\,\text{ESE},\ \overline{x} + 2.58\,\text{ESE}).$$

As with the z-test, these formulas should only be used if the sample size is at least 25.

And here's the alternative formula . . .

Example 12 Confidence interval for a population mean

Suppose a sample of size 30 has a sample mean of 40.3 and a sample standard deviation of 2.3. Then

$$n = 30, \quad \overline{x} = 40.3 \quad \text{and} \quad s = 2.3.$$

We have ESE $= 2.3/\sqrt{30} \simeq 0.4199$.

A 95% confidence interval for the population mean is

$$\begin{aligned}&(\overline{x} - 1.96\,\text{ESE},\ \overline{x} + 1.96\,\text{ESE})\\ &\quad \simeq (40.3 - 1.96 \times 0.4199, 40.3 + 1.96 \times 0.4199)\\ &\quad \simeq (40.3 - 0.8230, 40.3 + 0.8230)\\ &\quad \simeq (39.5, 41.1).\end{aligned}$$

Similarly, a 99% confidence interval for the population mean is

$$\begin{aligned}&(\overline{x} - 2.58\,\text{ESE},\ \overline{x} + 2.58\,\text{ESE})\\ &\quad \simeq (40.3 - 2.58 \times 0.4199, 40.3 + 2.58 \times 0.4199)\\ &\quad \simeq (40.3 - 1.0833, 40.3 + 1.0833)\\ &\quad \simeq (39.2, 41.4).\end{aligned}$$

We round the numbers forming the confidence interval to the same level of accuracy as the sample mean.

Activity 17 Calculating a confidence interval for a population mean

Suppose a sample of size 50 has a sample mean of 15.62 and a sample standard deviation of 6.44.

(a) Calculate a 95% confidence interval for the population mean.

(b) Calculate a 99% confidence interval for the population mean.

Activity 18 Calculating confidence intervals for jam

For a sample of 37 jars of a particular manufacturer's plum jam, the sum of the weights (in grams) and the sum of the squares of the weights were as follows:

$$\sum x = 16\,946, \quad \sum x^2 = 7\,762\,644.$$

(a) Calculate the mean and standard deviation of the weights in this sample.

(b) Calculate a 95% confidence interval for the mean weight of jars of plum jam produced by this manufacturer.

Activity 19 Which confidence interval is wider?

If both a 95% confidence interval and a 99% confidence interval were calculated for a population mean, which would be wider?

The interpretation of a confidence interval

We now have a way of finding a 95% confidence interval from a random sample, but we have not yet discussed how a confidence interval can be interpreted. Suppose we draw a very large number of different samples from some population with an unknown mean. Then although we do not know the value of the mean, from each sample we could form a confidence interval for it. Many of these confidence intervals will contain the population mean, but occasionally we will pick a sample that is not very representative of the population, and then the confidence interval might fail to contain it. For example, if we picked a random sample of twenty-five adults, we might by chance pick 24 men and only one woman. Then a confidence interval for, say, the population mean height might fail to contain the true mean height of all adults.

As a more specific example, suppose we pick a large number of samples of English state schools, each sample containing 100 schools. For each school we record the GCSE headline figure (P_{KS4}). We then calculate the mean and standard deviation of the 100 P_{KS4} values we obtained in each sample. The following are what these statistics might look like for, say, the first eight samples.

- Sample 1: $\overline{x} = 53.0$, $s = 16.32$
- Sample 2: $\overline{x} = 49.6$, $s = 17.36$
- Sample 3: $\overline{x} = 49.0$, $s = 14.82$
- Sample 4: $\overline{x} = 51.9$, $s = 16.81$
- Sample 5: $\overline{x} = 50.4$, $s = 15.51$
- Sample 6: $\overline{x} = 52.1$, $s = 17.55$
- Sample 7: $\overline{x} = 54.8$, $s = 16.68$
- Sample 8: $\overline{x} = 50.9$, $s = 17.94$

For each sample, we can calculate a 95% confidence interval for the population mean. The formula for the interval is:

$$(\overline{x} - 1.96s/\sqrt{100}, \overline{x} + 1.96s/\sqrt{100}).$$

For the above eight samples, the 95% confidence intervals are:

- Sample 1: $(49.8\%, 56.2\%)$
- Sample 2: $(46.2\%, 53.0\%)$
- Sample 3: $(46.1\%, 51.9\%)$
- Sample 4: $(48.6\%, 55.2\%)$
- Sample 5: $(47.4\%, 53.4\%)$
- Sample 6: $(48.7\%, 55.5\%)$
- Sample 7: $(51.5\%, 58.1\%)$
- Sample 8: $(47.4\%, 54.4\%)$

Suppose, now, that the mean GCSE headline figure for English state schools is actually 51.0%. Then all but one of the above confidence intervals contain the true population mean. (The exception is Sample 7.)

In fact, if we had a very large number of samples, and calculated a 95% confidence interval from each of them, 95% of the confidence intervals would contain the population mean.

This leads to the following interpretation of a 95% confidence interval.

Confidence intervals and the population mean

About 95% of the possible random samples we could select will give rise to a 95% confidence interval that *does* include the population mean.

Of the remaining intervals, half of them will be completely below the population mean, and the other half will be completely above the population mean. That is, about 2.5% will give intervals that are completely below the population mean and about 2.5% will give intervals completely above the population mean.

Thus, about 5% of possible random samples that might be selected will give rise to a 95% confidence interval that *does not* include the population mean.

So if you say that a 95% confidence interval includes the population mean, you will be right 95% of the time; that is, you can be 95% confident that your statement is correct.

Another way of thinking of a confidence interval is from the point of view of a working statistician who calculates 95% confidence intervals on many occasions. On about 95% of the occasions the interval will include the mean, but on the other 5% of occasions it will not.

An important point to note is the very close link between a hypothesis test and a confidence interval. A confidence interval contains the values that the population mean might plausibly equal, and the purpose of a hypothesis test is to reject implausible values. From the way we obtain the formula for a confidence interval, we have an exact link.

Confidence intervals and hypothesis tests

If the confidence interval does include the hypothesised population mean, we do not reject the hypothesis. If the confidence interval does not include the hypothesised population mean, then we do reject the hypothesis. In particular:

- If the 95% confidence interval does not include the hypothesised population mean, then we reject the hypothesis at the 5% significance level.
- If the 99% confidence interval does not include the hypothesised population mean, then we reject the hypothesis at the 1% significance level.

Activity 20 Doing hypothesis tests about jam

In Activity 18, you calculated a 95% confidence interval for the mean weight of jars of plum jam produced by a particular manufacturer. Let μ be this mean weight. On the basis of your confidence interval, what can you say about the conclusion that would be drawn from each of the following hypothesis tests?

(a) A hypothesis test of $H_0:\ \mu = 454$ against $H_1:\ \mu \neq 454$.

(b) A hypothesis test of $H_0:\ \mu = 457$ against $H_1:\ \mu \neq 457$.

You have now covered the material related to Screencasts 4 and 5 for Unit 9 (see the M140 website).

4.3 An interval estimate for the difference between two means

We can also derive a confidence interval from two-sample z-tests (introduced in Section 6 of Unit 7). These tests are used when we have one sample from a population with mean μ_A and a second sample from a population with mean μ_B. The hypothesis test examines whether these population means might be equal. The hypothesis test requires the two sample sizes to both be 25 or more. Provided this condition is satisfied, then the samples can also be used to form a confidence interval for the difference between the two population means, $\mu_A - \mu_B$.

As before, we obtain the confidence interval by trying various values of the quantity in which we are interested: in this case, $\mu_A - \mu_B$. As the sample sizes are large, we know that

$$z = \frac{(\overline{x}_A - \overline{x}_B) - (\mu_A - \mu_B)}{\text{ESE}}$$

has approximately the standard normal distribution. The method for obtaining a confidence interval is to calculate z for different values of $\mu_A - \mu_B$ and carry out a hypothesis test each time. If the null hypothesis is rejected for a particular value of $\mu_A - \mu_B$, then that value is not included in the confidence interval; otherwise it is included. This gives the following procedure.

Calculating 95% and 99% confidence intervals for $\mu_A - \mu_B$, the difference between two population means

Suppose the two sample sizes are n_A and n_B, the sample means are $\overline{x}_A$ and $\overline{x}_B$, and the sample standard deviations are s_A and s_B.

- Calculate the estimated standard error:

$$\text{ESE} = \sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}.$$

- The 95% confidence interval for $\mu_A - \mu_B$ is

$$(\overline{x}_A - \overline{x}_B - 1.96\,\text{ESE},\ \overline{x}_A - \overline{x}_B + 1.96\,\text{ESE}).$$

- The 99% confidence interval for $\mu_A - \mu_B$ is

$$(\overline{x}_A - \overline{x}_B - 2.58\,\text{ESE},\ \overline{x}_A - \overline{x}_B + 2.58\,\text{ESE}).$$

These formulas should only be used if the sample sizes are at least 25.

Example 13 99% confidence interval estimate for $\mu_A - \mu_B$

Activity 6 (Subsection 1.2) concerned the GCSE headline figures (P_{KS4}) in community schools compared with other schools. The data that were used to make the comparison were given in Table 2, reproduced here as Table 7.

Table 7 Summary statistics for P_{KS4} by type of school

	Sample size	Sample mean (%)	Sample standard deviation (%)
Community school	43	49.8	13.55
Other school	57	50.7	19.61

We will calculate a 99% confidence interval for $\mu_A - \mu_B$, where μ_A is the average GCSE headline figure in community schools, while μ_B is the average in other schools. We first calculate the ESE:

$$\text{ESE} = \sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}} = \sqrt{\frac{13.55^2}{43} + \frac{19.61^2}{57}} \simeq 3.319.$$

Now

$$\begin{aligned}
&(\overline{x}_A - \overline{x}_B - 2.58\,\text{ESE}, \overline{x}_A - \overline{x}_B + 2.58\,\text{ESE})\\
&\quad \simeq (49.8 - 50.7 - 2.58 \times 3.319, 49.8 - 50.7 + 2.58 \times 3.319)\\
&\quad \simeq (-0.9 - 8.56, -0.9 + 8.56)\\
&\quad \simeq (-9.5, 7.7).
\end{aligned}$$

So the 99% confidence interval for $\mu_A - \mu_B$ is $(-9.5\%, 7.7\%)$. Notice this includes zero. This was expected, as in Activity 6 we did not reject the null hypothesis that the difference between the population means was zero, so zero is a plausible value for the difference.

Activity 21 95% confidence interval estimate for $\mu_A - \mu_B$

Calculate the 95% confidence interval for the difference between the average GCSE headline figure in community schools and the average GCSE headline figure in other schools.

Activity 22 Investigating the effect of siblings

A researcher set out to investigate children's verbal skills at the age of six. She wished to compare the performance of only children (which she defined as children with no brothers or sisters, and not living in a household with other children) with that of children who live with at least one other child.

The researcher devised a short test, and gave it to two random samples of six-year-old children: one sample of only children and one sample of children, 'other' children, who live with at least one other child. The results are summarised in the following table.

	Sample size	Sample mean	Sample standard deviation
Only children (A)	100	58.4	18.2
Other children (B)	150	52.1	20.6

Calculate a 95% confidence interval for the difference in average scores between only children and other children.

Suppose a hypothesis test were performed to compare the average scores of these two groups of children. On the basis of the confidence interval, what can be said about the conclusion that would be made from the test?

Exercises on Section 4

Exercise 9 Quantifying the size of skulls

In an investigation of skull size, the breadth of 30 male Egyptian skulls dating from around 4000 BC were measured. The mean and standard deviation of these measurements (in millimetres) were 131.4 and 5.1, respectively.

(a) Calculate a 95% confidence interval for the breadth of male Egyptian skulls of that period.

(b) If a 99% confidence interval were calculated, would it be wider or narrower than the interval that you calculated in (a)?

Exercise 10 Measuring the changing size of skulls

As part of the investigation described in Exercise 9, the breadth of 30 male Egyptian skulls from about 1850 BC were also measured. The sample mean was 134.5 mm, and the sample standard deviation was 3.4 mm.

Calculate a 95% confidence interval for the change in mean breadth of male Egyptian skulls from 4000 BC to 1850 BC. Has the mean breadth changed over that time?

5 Interval estimates from fitted lines

In the previous section you learned about interval estimates. In particular, you learned how to calculate and interpret confidence intervals for means. In this section you will learn about interval estimates in another context – least squares regression.

Recall that in Section 4 of Unit 5 we found that a linear relationship can be modelled by fitting a least squares regression line to the data. The line can be written in the form of the equation

$$y = a + bx,$$

where a is the intercept of the line and b is the slope of the line. The regression line can then be used to estimate the value of the dependent variable for a known value of the explanatory variable. The fitted value $y = a + bx$ is an estimate of the average value of y when the explanatory variable has the value x.

So least squares regression lines can be used to make predictions. This is important, and leads to the following two types of interval estimate:

- The *confidence interval for the mean response*, which provides an interval for the position of the regression line. This is introduced in Subsection 5.1.
- The *prediction interval*, which provides an interval for the prediction of a new value. This is introduced in Subsection 5.2.

In Subsection 5.3, you will be referred to the Computer Book to explore confidence intervals further and learn how to produce these confidence intervals and prediction intervals using Minitab.

5.1 Confidence intervals for the mean response

In Subsection 3.1, Example 7 introduced some data on the pass rates in English and Mathematics. Suppose that y represents the pass rate in Mathematics and x represents the pass rate in English. Then the least squares fitted line for the data in Figure 11 of the example has the following equation:

$$y = 15.9 + 0.720x.$$

This line can be used to predict the pass rate in Mathematics when the pass rate in English is known. For example, for schools where the pass rate in English is 50%, the pass rate in Mathematics is estimated to be

$$15.9\% + 0.720 \times 50\% = 51.9\%.$$

This estimate of 51.9% is a single number, so it is a point estimate. Although a point estimate may be a 'best guess', it says nothing about what other values are also plausible. For this, an interval estimate is required.

In Section 4 you saw that a confidence interval for a mean can be calculated by thinking about the range of values which would not be rejected in a z-test. In this section we will take a slightly different, but related, approach to thinking about interval estimates – we will consider what might have happened if we had taken another sample. To do this we need to introduce the concept of a **population least squares regression line**. This is the regression line that we would calculate if we knew all values in the whole population for both variables, instead of just those in the sample we have observed.

We shall write the equation of the population least squares regression line as $y = \alpha + \beta x$. The quantity β is then the **population slope of the least squares regression line**, and α is the **population intercept of the least squares regression line**. The coefficients of the least squares regression line calculated from a sample are then just sample estimates of α and β. It is to be expected that the sample slope and the sample intercept will be close to these population values – but how close?

α and β are the first two letters of the Greek alphabet. α is 'alpha' and β is 'beta'.

One way of exploring this is to think about what would happen if we took a number of separate samples from the population and calculated a least squares regression line for each sample. For example, suppose a population consists of 500 individuals, and the observations of two variables x and y on all 500 individuals were as follows.

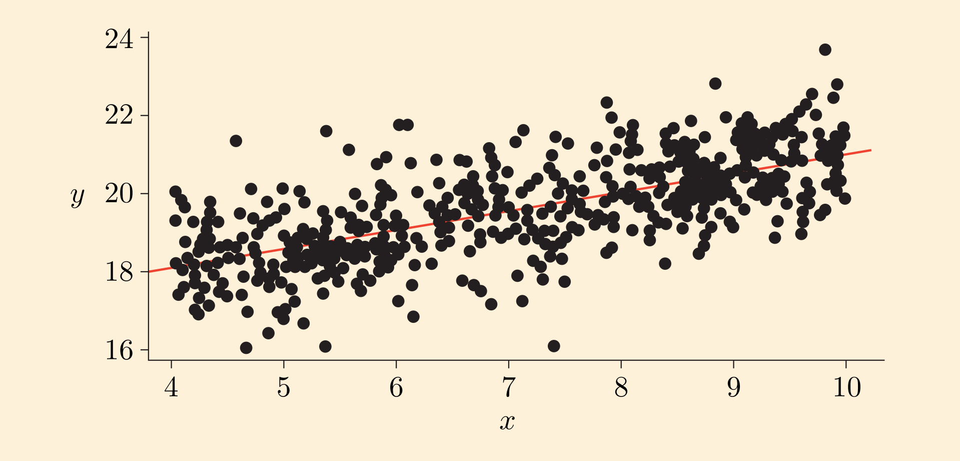


Figure 27 A scatterplot of data for a population of 500 individuals

For this population the relationship between x and y is positive, linear and reasonably strong. The least squares regression line summarising these data has the equation

$$y = 16.15 + 0.48x.$$

Now suppose a random sample of 20 individuals is taken from the population and the regression line given by the sample is calculated. One such sample, and the resulting regression line, is shown in Figure 28.

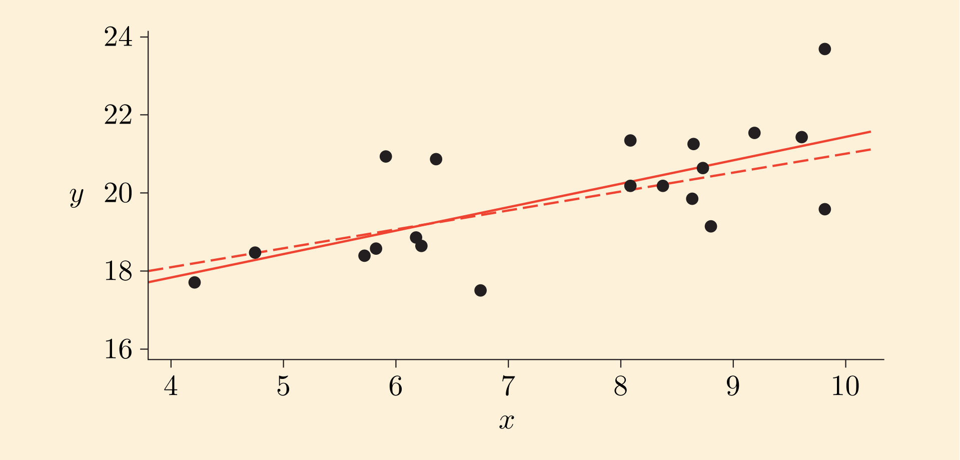


Figure 28 A scatterplot of data for a sample of 20 individuals. The solid line gives the least squares regression line based on the sample. The dashed line gives the least squares regression line for the population

The equation of the regression line for this sample is

$$y = 15.42 + 0.603x.$$

So for this sample, the intercept happens to be a bit lower than the correct value for the whole population, and the slope turns out to be a bit higher. Overall, the regression line for the sample is below the population line for values of x that are less than 6, and above the population line for x-values that are greater than 6.

If other samples are taken, the position of the regression line will differ from sample to sample, fluctuating around the regression line given by the whole population. This is illustrated in Figure 29, where regression lines from 100 different random samples (each of size 20) are shown.

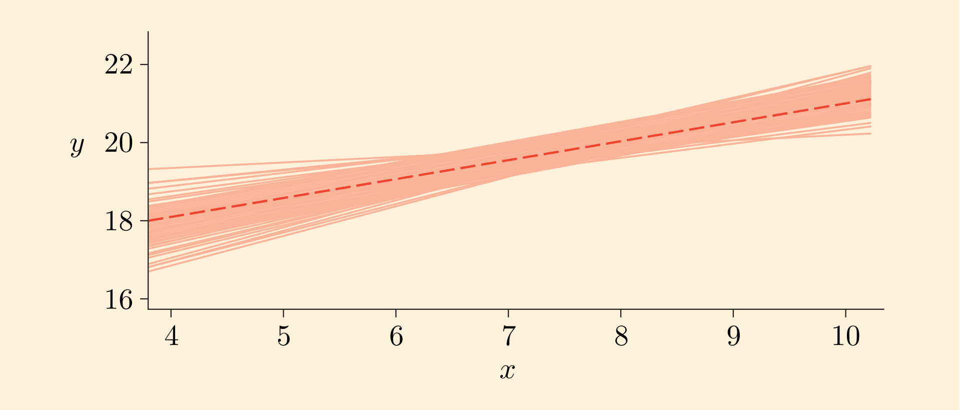


Figure 29 Regression lines calculated using 100 different samples of 20 individuals. The solid lines are the regression lines from the samples. The dashed line gives the regression line for the population

Notice that the question of which sample lines are most extreme depends on the value of x being considered. Different lines are the most extreme ones for different values of x.

To clarify the variability in the lines, we could plot, for every value of x, the band within which 95% of the sample lines lie. This is done, in Figure 30, for the population of 500 individuals shown in Figure 27. The band is wider towards the ends of the range of x than at the centre of the range, and symmetrical about the population regression line.

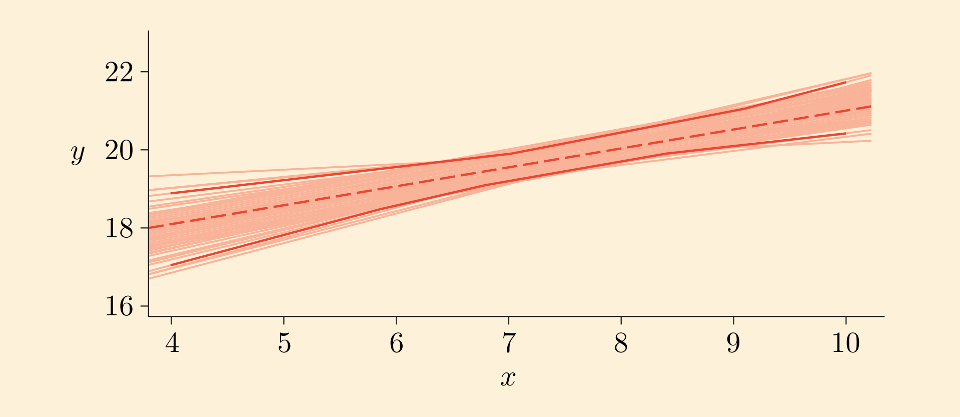
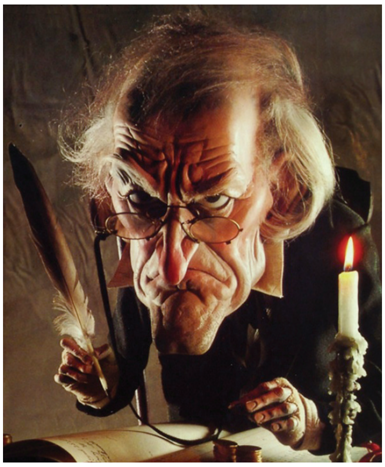


Figure 30 Least squares regression lines calculated using 100 different samples of 20 individuals. The solid curves delineate the band within which 95% of the sample least squares regression lines lie for each value of x. The dashed line gives the least squares regression line for the population

Figure 30 characterises the differences between the population regression line and the sample regression lines. Usually, of course, we only have the regression line from one sample, rather than a regression line from each of 100 samples. We would like to use that sample regression line to estimate the plausible points that the population regression line might pass through.

How might this be achieved? The approach usually taken is to make assumptions about how points are scattered around the population line. This enables statements to be made about the population line using just one sample of points. In particular, for any value of the explanatory variable x, it allows a **confidence interval for the mean value of** y (the 'mean response') to be constructed.

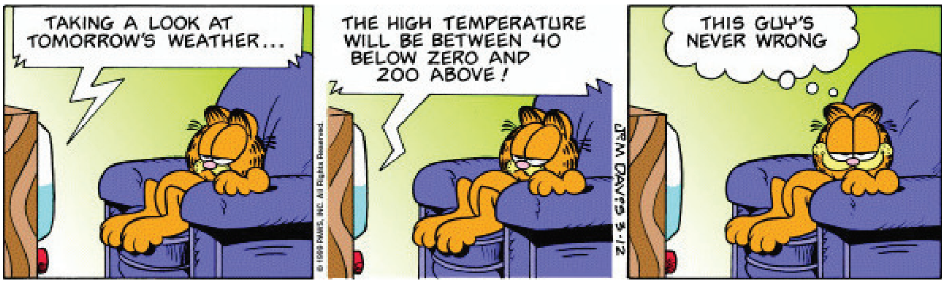


Dickens's Scrooge: someone capable of a mean response!

Now suppose we are given a confidence interval for the mean response for a particular value of x, the interval $(y_{\text{min}}, y_{\text{max}})$. If we happened to know the population slope and intercept, we could also calculate what the 'correct' answer is, say y_{true}. Then our confidence interval either contains y_{true} or it does not. And it would be helpful if it did.

(In practice we do not know the population slope and intercept. If we did, there would be no point working with the sample values.)

It is not possible to construct a useful interval from sample data that is guaranteed to always contain y_{true}. Instead the confidence interval $(y_{\text{min}}, y_{\text{max}})$ is constructed in such a way that there is a high chance it will contain y_{true}. Typically this 'high chance' is taken to be a probability of 95%, so the 95% confidence interval for the mean response is formed. That is, with 100 such intervals, on average 95 of them will contain y_{true}. So given just one interval, saying that it contains y_{true} is the right thing to say 95% of the time.


TAKING A LOOK AT TOMORROW'S WEATHER...
THE HIGH TEMPERATURE WILL BE BETWEEN 40 BELOW ZERO AND 200 ABOVE!
THIS GUY'S NEVER WRONG
© 1998 PAWS, INC. All Rights Reserved.
JIM DAVIS 3-12

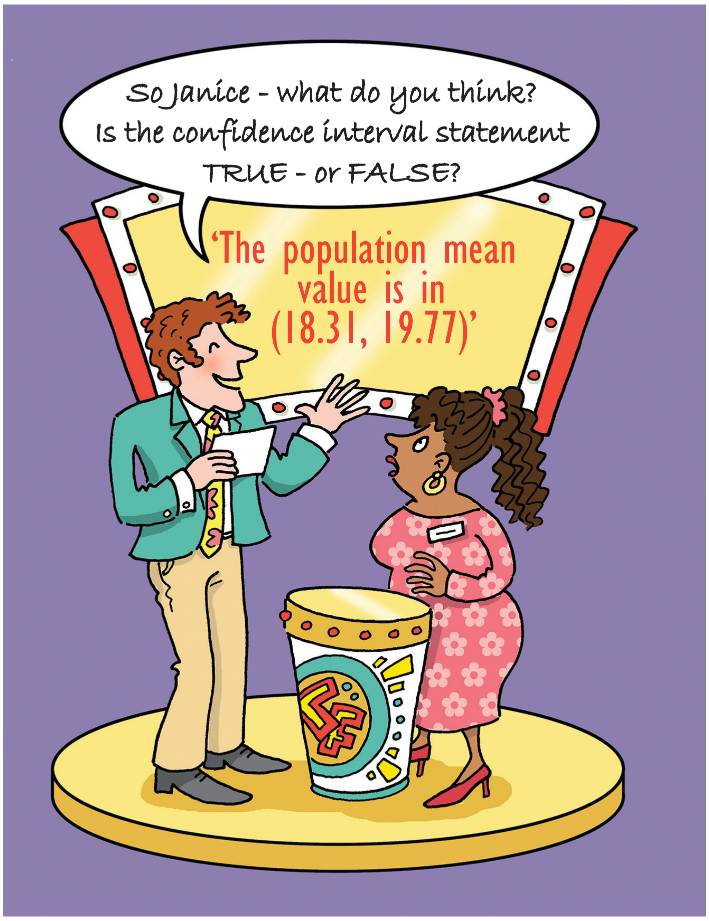
Example 14 Interpreting a confidence interval for the mean response

Returning to the sample plotted in Figure 28, the 95% confidence interval for the mean response when $x = 8$ turns out to be (19.67, 20.82).

This means that there is a 95% chance that the statement 'The interval (19.67, 20.82) contains the (population) mean value of y when $x = 8$' is true.

Activity 23 Interpreting another confidence interval

Using the population line, calculate the population mean of y when $x = 6$. Is the statement 'The interval (18.31, 19.77) contains the (population) mean value of y when $x = 6$' true or false?


So Janice - what do you think? Is the confidence interval statement TRUE - or FALSE?
'The population mean value is in (18.31, 19.77)'

Activity 24 Interpreting yet another confidence interval

From a different population, 100 samples were taken. For each sample, the confidence interval for the mean response was calculated for $x = 7$. These confidence intervals are plotted in Figure 31. The population least squares line was also determined and gives a value $y = 18.55$ when $x = 7$. The line corresponding to this value of y is also plotted. In Figure 31 how many of the confidence intervals do not contain the population mean value of y? Hence what is the probability that a randomly chosen confidence interval out of these 100 intervals does include the population mean value?

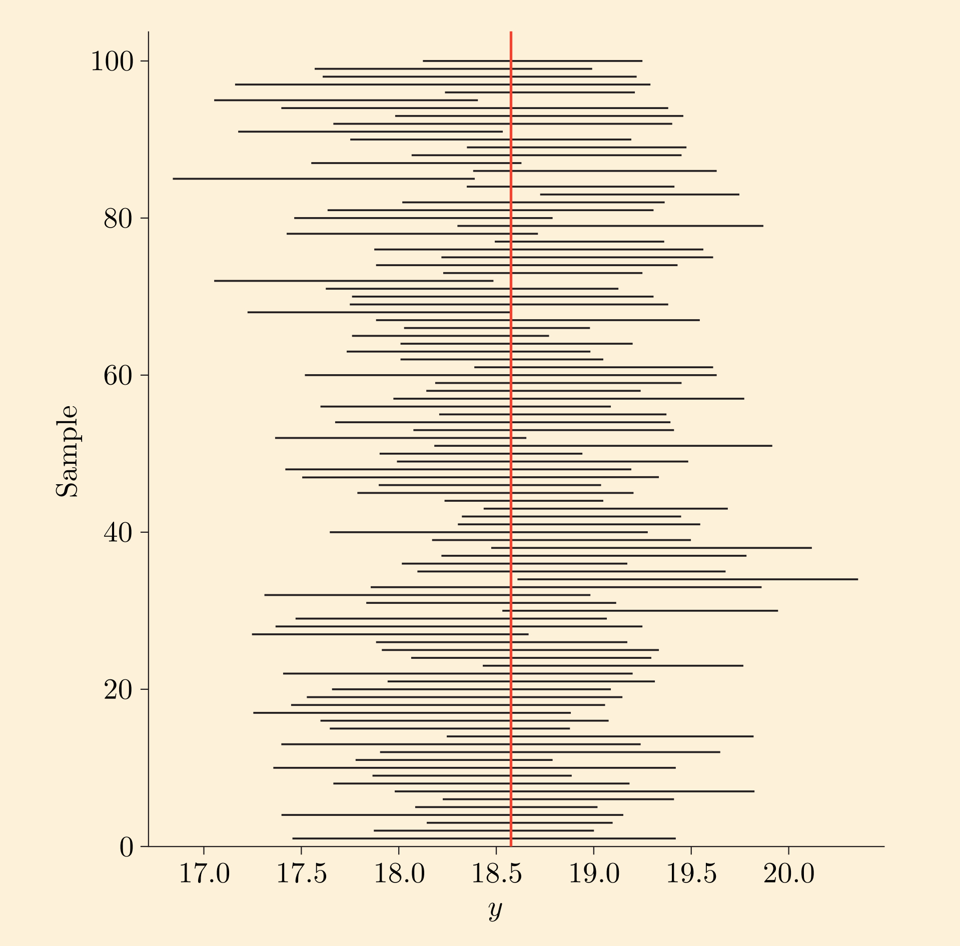


Figure 31 Confidence intervals for the mean response when $x = 7$ for 100 different samples

Activity 25 Interpreting a confidence interval for a mean pass rate

For schools with a pass rate in English of 50%, the predicted pass rate in Mathematics is 51.9%. The 95% confidence interval for the mean response turns out to be 48.6% to 55.2%. Interpret this confidence interval.

For any sample, it is possible to map the resulting confidence intervals for the mean response for a range of different x-values. For example, in Figure 32 the confidence intervals for a range of different values of x are plotted. Notice that these confidence intervals have different widths, narrower in the centre, and wider as you move away from the centre.

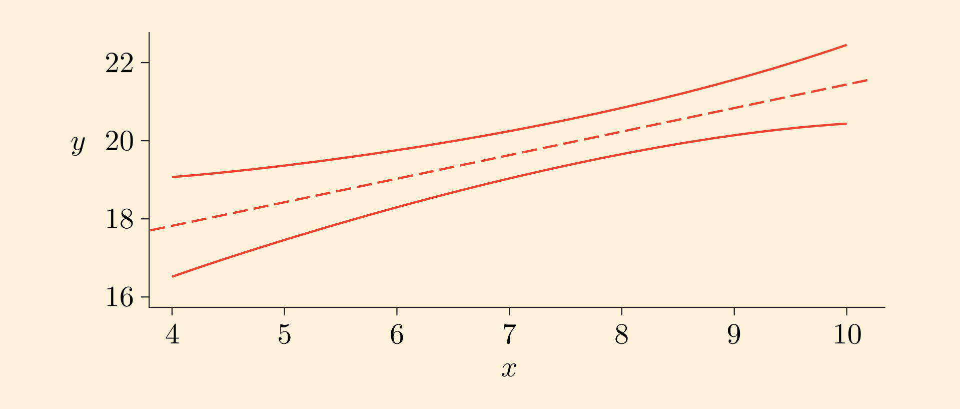


Figure 32 Limits of the 95% confidence intervals for the mean response (solid curves) based on one sample. The dotted line gives the least squares regression line based on the same sample

This pattern mirrors what you saw in Figure 30 and reflects two properties of confidence intervals for the mean response, as follows.

Properties of the confidence interval for the mean response

1. Confidence intervals for the mean response are at their narrowest when x is the sample mean, and get steadily wider each side.
2. The least squares regression line based on the sample data is always in the middle of the interval. Hence the point estimate for the predicted value is always in the middle of the confidence interval for the mean response.

Example 15 Deducing a point estimate from a confidence interval

In Activity 25, a 95% confidence interval for the mean pass rate in Mathematics was 48.6% to 55.2% for schools with a pass rate in English of 50%.

The midpoint of this interval is $(48.6\% + 55.2\%)/2 = 51.9\%$. This is the same as the predicted pass rate in Mathematics for such schools.

Activity 26 Comparing widths of confidence intervals

The width of the confidence interval given in Activity 25 is $55.2\% - 48.6\% = 6.6\%$. Confidence intervals for various pass rates in English are given in Table 8.

Table 8 Some confidence intervals for the mean pass rate in Mathematics

Pass rate in English	Confidence interval for mean pass rate in Mathematics	Width of confidence interval
40	(40.1%, 49.3%)	
50	(48.6%, 55.2%)	6.6%
60	(56.9%, 61.3%)	
70	(64.5%, 68.1%)	
80	(71.0%, 75.9%)	

(a) Complete the final column of Table 8.

(b) From Table 8, which confidence interval is narrowest? State the average pass rate in English to the nearest 10%.

5.2 Prediction intervals

In Subsection 5.1, interval estimates bracketing the position of the population line were introduced. However, for a given value of the explanatory variable, these intervals only indicate a range for the population mean – not a range for an individual value. Even if we know the exact position of the population line, that will not allow us to pinpoint where an individual point lies. Some individuals will be above the population line, whilst others will be below the population line. So for predictions about individuals, a different interval is required – a so-called **prediction interval**.

Example 16 Prediction of a pass rate for a particular school

In Activity 25 (Subsection 5.1) the confidence interval for the mean pass rate in Mathematics for schools with a pass rate of 50% in English was given as 48.6% to 55.2%.

Now, not all schools with a pass rate in English of 50% will have the same pass rate in Mathematics. Some will have a higher pass rate than the average, whilst others will not have quite as good a pass rate. So the interval 48.6% to 55.2% does not give a sufficient reflection of the variability in Mathematics pass rates between individual schools. A prediction interval takes such variability into account. For schools with a pass rate in English of 50%, the 95% prediction interval turns out to be 33.6% to 70.3%. So whilst on average such schools will have a pass rate in Mathematics of about 50%, for some schools the pass rate will be much higher, and for others it will be much lower.

Prediction intervals

Prediction intervals reflect the random variation of individual values around the population regression line, as well as uncertainty about the actual position of that line. If we have a very large sample, then we will have a good idea about the position of the population regression line – most of the uncertainty in predicting the value of an individual will stem from the scatter of individual values about the regression line.

95% prediction intervals are calculated so that 95% of them will contain the actual value a new observation would take.

Properties of prediction intervals

Prediction intervals inherit many of the properties of the confidence intervals for the mean response. In particular, a prediction interval:

- is centred around the predicted value $a + bx$
- is narrowest when x is the sample mean, and steadily widens away from this point
- gets wider as the scatter around the line increases.

Also, a prediction interval is always wider than the corresponding confidence interval for the mean response.

Activity 27 Valid prediction interval?

For a school with a pass rate in English of 35%, the 95% confidence interval for the mean response is 35.9% to 46.4%. Two different students quote the corresponding prediction interval. Ali says the interval is 37.8% to 44.5%, and Charlie says the interval is 17.3% to 55.0%.

Explain why both Ali and Charlie must be mistaken.

You have now covered the material related to Screencast 6 for Unit 9 (see the M140 website).

5.3 Computer work: interval estimates

In Sections 4 and 5 you have been introduced to various interval estimates. Although you have learned how to calculate some of these intervals, the emphasis has been on the interpretation of the intervals. In this subsection you will learn how to use Minitab to obtain various interval estimates. You will learn how to produce confidence intervals for the mean based on the one-sample z-test, and how to obtain confidence intervals for the mean response and prediction intervals. You will also further explore the confidence intervals for the mean response.

In M140, you will not be expected to calculate confidence intervals for the mean response or prediction intervals by hand.

You should now turn to the Computer Book and work through Subsection 9.1 if you have not already done so, followed by the rest of Chapter 9.

Exercises on Section 5

Exercise 11 Predicting the effect of a drug

In Subsection 3.3 of Unit 5, some data on the effect of a drug, captopril, were introduced. The data consisted of diastolic blood pressure measurements before, and two hours after, injection with captopril for 15 patients. Using those data the following least squares line was calculated:

$$y = 4.2 + 0.880x,$$

where y is a patient's diastolic blood pressure two hours after injection with captopril, and x is the patient's diastolic blood pressure before the injection.

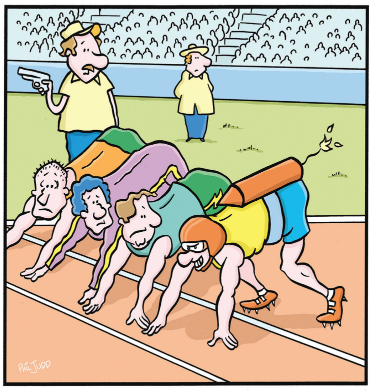
(a) Using the least squares regression line, calculate the diastolic blood pressure two hours after the injection for patients with an initial diastolic blood pressure of 110 mmHg. Why might an interval estimate be preferred?

(b) The 95% confidence interval for the mean post-injection diastolic blood pressure is 95.9 to 106.1 mmHg. Does this suggest that captopril is effective in reducing blood pressure in patients with an initial diastolic blood pressure of 110 mmHg? (Here, take 'effective' to mean that the diastolic blood pressure is, on average, reduced.)

(c) Does the confidence interval given in part (b) suggest that for any particular patient with an initial diastolic blood pressure of 110 mmHg, captopril is likely to be effective?

Exercise 12 Predicting performance in sport

Exercise 3 (Section 2) showed the best times for the 200-metre and 100-metre sprints by UK male sprinters in 2011.



A few UK male sprinters only had a time recorded for the 200-metre sprint. Using a least squares regression line, it is possible to predict a time for the 100-metre sprint that these sprinters might have achieved in 2011.

(a) For one such sprinter, the 95% prediction interval for his best 100-metre time in 2011 is 9.88 seconds to 10.62 seconds. Interpret the meaning of this interval.

(b) What must have been the point estimate associated with the prediction interval given in part (a)?

(c) The 'B' standard qualifying time for the men's 100 metres at the 2012 Olympics was 10.24 seconds. Is it plausible that the sprinter in part (a) could have achieved this qualifying time? Why or why not?

(d) Figure 33 is a scatterplot of the best times for the 200-metre sprint and 100-metre sprint achieved in 2011 for 100 UK sprinters, along with the 95% prediction intervals. On the plot, the points that lie below the 95% prediction interval and the points that lie above the 95% prediction interval are marked with different symbols. Based on this, do the prediction intervals look reasonable? Justify your answer.

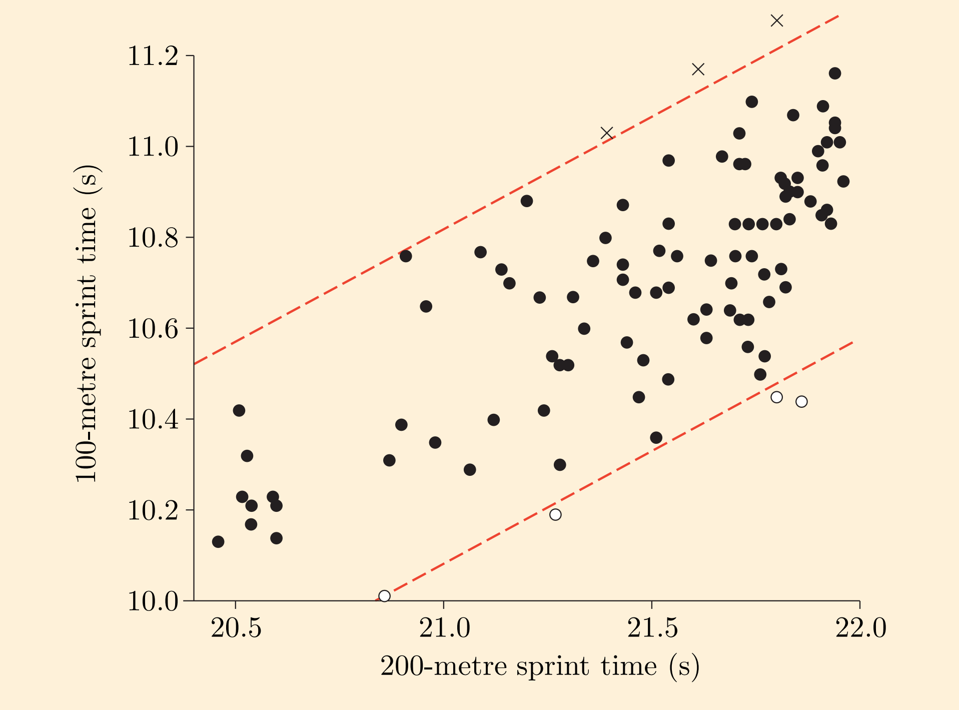


Figure 33 Times for male UK sprinters in 2011 with 95% prediction intervals

6 Case study

In Section 1, we began measuring the quality of an English secondary school using the GCSE headline figure (P_{KS4}) – the percentage of students at the end of

Key Stage 4 who achieve at least five A* to C grade GCSEs including English and Mathematics. Performance in GCSEs relates both to characteristics of the school and characteristics of the students. A school that attracts academically able students and then teaches them poorly could have a higher GCSE headline figure than a school that attracts academically less able students and then teaches them well. Hence, in assessing a school's performance, some account should perhaps be taken of students' academic ability when they started at the school. The purpose of this case study is to explore one possible way of doing this.

In Subsection 1.2, a measure of the academy ability of eleven-year-old students was introduced: P_{KS2} – the average points score of students at the end of Key Stage 2. As the end of Key Stage 2 is the end of primary education, P_{KS2} only indicates the ability of a secondary school's intake, not any impact of the secondary school's teaching. We are focusing on the P_{KS4} of 100 secondary schools in 2011. The cohort of students taking GCSEs in 2011 took Key Stage 2 tests in 2006. The Key Stage 2 average point score of a school's intake in 2006 will be referred to as its P_{KS2}. As with P_{KS4}, a higher value of P_{KS2} indicates that the students were on average doing better in national examinations. Here we consider using P_{KS2} to predict P_{KS4}, and examine whether taking account of P_{KS2} would change perception of which schools perform well.

Activity 28 Describing the relationship between P_{KS2} and P_{KS4}

A scatterplot of the Key Stage 2 results and the Key Stage 4 results for each of the 100 schools is given in Figure 34.

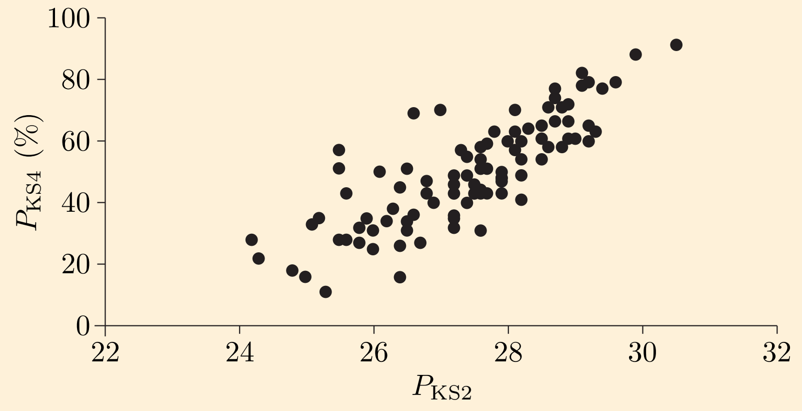


Figure 34 Scatterplot of P_{KS4} and P_{KS2}

(a) Describe the relationship between P_{KS4} and P_{KS2}.

(b) Is the correlation coefficient likely to be a good guide to the strength of the relationship between P_{KS4} and P_{KS2}? Why or why not?

(c) Are there any outliers or influential points? Justify your answer.

(d) The correlation coefficient turns out to be $+0.83$. Does this make sense in the light of your answer to part (a)?

As noted in Activity 28, the relationship between P_{KS2} and P_{KS4} appears to be reasonably linear. So the relationship between the two variables can be summarised by a straight line with P_{KS4} as the response variable and P_{KS2} as the explanatory variable.

Activity 29 Modelling Key Stage 4 results

Calculate the equation of the least squares regression line using the data from the 100 non-selective schools depicted in Figure 34, giving the slope to three decimal places and the intercept to two decimal places. (The calculation of the least squares regression line was introduced in Subsection 4.2 of Unit 5.) To reduce the amount of computation involved, the first step of the calculation has already been completed for you:

$$\sum x = 2743.6, \quad \sum y = 5031,$$
$$\sum x^2 = 75\,449.44, \quad \sum y^2 = 282\,359, \quad \sum xy = 139\,903.0.$$

The least squares regression line, along with the 95% confidence interval for the mean response, is shown in Figure 35.

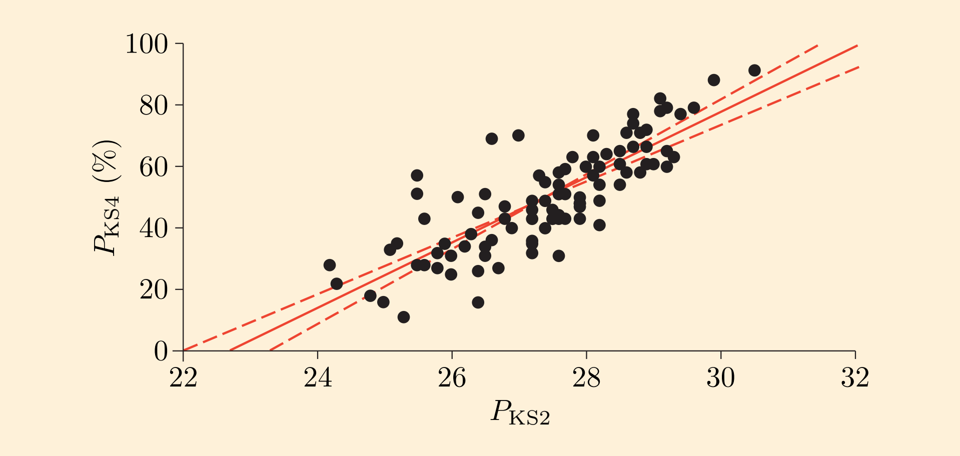


Figure 35 Least squares regression line and 95% confidence interval for the regression of P_{KS4} on P_{KS2}

Activity 30 Interpreting the model

Using Figure 35, give ranges for the mean GCSE headline figure for the following values of P_{KS2}. (As you are reading off a graph, all the numbers are expected to be approximate.)

(a) 24

(b) 28

(c) 30

From Figure 35, notice that values between the confidence limits all indicate a positive relationship between Key Stage 2 and Key Stage 4 results.

Activity 31 Measuring the relationship between Key Stage results

Use the values given in Activity 29 to verify the value of the correlation coefficient between the Key Stage 2 and Key Stage 4 results given in Activity 28.

Activity 32 Comparing schools – 1

Table 9 gives the GCSE headline figure and the Key Stage 2 average point score for each of three schools.

Table 9 Results for three schools

School	Key Stage 2	Key Stage 4
A	29.3	63%
B	28.1	57%
C	25.5	51%

(a) Using the least squares regression line that you calculated in Activity 29, predict the results at Key Stage 4 for each school on the basis of their Key Stage 2 results. (Give your answers to one decimal place.)

Residuals were introduced in Subsection 3.2 of Unit 5.

(b) For each school, calculate the residual. That is, calculate the residuals between the actual Key Stage 4 results and the predicted ones. Rank the schools on the basis of these residuals. Hence comment on the relative quality of education at the schools, as measured by the Key Stage results.

As you have seen in the last couple of activities, a least squares regression line can be used to predict the results a school will get at Key Stage 4 given the ability of its intake at Key Stage 2. However, as you learned in Section 5, by itself a point estimate conveys insufficient information. Also important is some indication of how accurate that estimate is.

For the three schools described in Activity 32 the prediction intervals are as follows.

Table 10 Prediction intervals for three schools

School	Key Stage 2	Key Stage 4	Prediction	Prediction interval
A	29.3	63%	70.1%	$(50.5\%, 89.8\%)$
B	28.1	57%	57.4%	$(37.9\%, 76.9\%)$
C	25.5	51%	29.7%	$(10.1\%, 49.4\%)$

Look at School A. As noted in Activity 32, 63% of its pupils achieved the Key Stage 4 benchmark in 2011, whilst a figure of just over 70% was expected on the basis of its intake. However, 63% is inside the prediction interval. So, on the basis of this model, just 63% achieving the benchmark is not exceptional.

'We were finding it impossible to get the kids into a decent school, so we had them adopted.'

Activity 33 Comparing schools – 2

Use the prediction intervals given in Table 10 to comment on the performance of schools B and C in 2011.

An overall picture of the prediction intervals is given in Figure 36. As in Figure 33 (in Exercises for Section 5), the schools whose GCSE headline figure lies outside the corresponding prediction interval are indicated using different plotting symbols. There were six such schools.

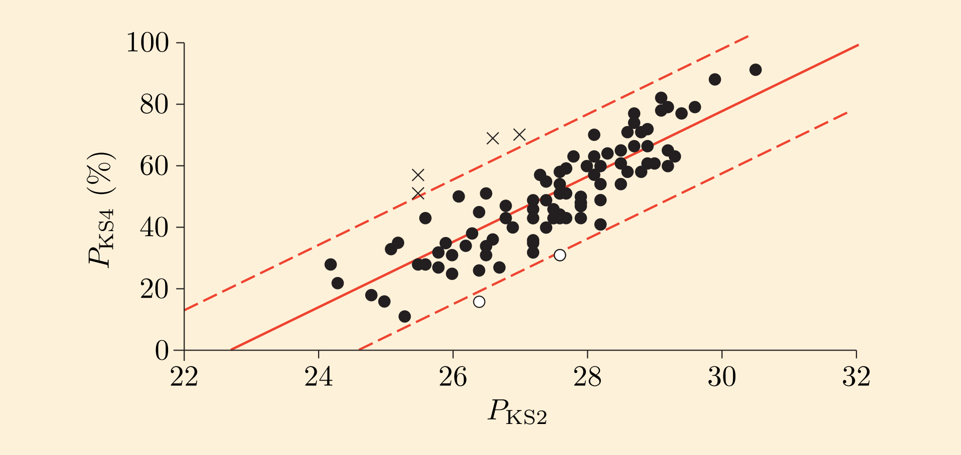


Figure 36 Least squares regression line and 95% prediction interval for the regression of P_{KS4} on P_{KS2}

Activity 34 Judging the model

Does Figure 36 indicate that the model is not correct?

Activity 35 Comparing school performance measures

Consider the six schools that are outside the 95% prediction interval in Figure 36. If the Key Stage 2 results are ignored, is their performance at Key

Stage 4 notable?

We have seen that the Key Stage 2 results can be used to predict the percentage of students who should get at least five A* to C grade GCSEs including English and Mathematics. Comparing this with the actual percentage gives a measure of the performance of the schools that allows for the ability of its students when they started at the school. Thus it provides a value-added measure of school performance that is an indication of how much the school has enhanced the ability of its students.

However, there are drawbacks to this approach. In this section we have been considering a line based on the data from just 100 secondary schools. Using information from more schools would improve the estimate. For example, a model could be fitted to all the data from 2011. But what about other years? A judgement would have to be taken as to whether it is reasonable to believe that the relationship between P_{KS2} and P_{KS4} is the same in other years as well.

We have only considered a linear relationship between P_{KS2} and P_{KS4}. From Figure 34 this assumption looks reasonable. However, other more complicated relationships might fit the data better and lead to different predictions. These differences are likely to be only small, but small differences can be important if great emphasis is placed on this measure of school performance. Furthermore, we have only looked at using P_{KS2} to predict P_{KS4}. Perhaps another measure is better. For example, ability at Key Stage 2 could be measured by just taking the results in English and Mathematics, and/or by considering the percentage who attain at least Level 4 (the expected standard at Key Stage 2) instead of the average point score. Also, should other factors about the school be taken into account – for example, the percentage of students from a disadvantaged background?

It must also be remembered that P_{KS4} is only one measure of secondary school performance. Even if exam results are seen as the be all and end all of school education, it is arguable whether P_{KS4} is the right measure. Should the range of grades be extended to A* to G instead of just A* to C? Should the range of qualifications include just English and Mathematics? Should it include more subjects? Changing the summary measure of attainment at Key Stage 4 will change the apparent performance of some schools.

Finally, with a lot potentially riding on a school's position relative to other schools, it is important that the means by which the performance is calculated is open, transparent and not subject to dispute. This requirement provides one explanation as to why league tables of school performance in 2011 chose only to focus on results at Key Stage 4, and not make any adjustment for the students' ability when they started at the school, though of course this might not be popular with schools who have a low-attaining intake.

Exercises on Section 6

Exercise 13 Predicting P_{KS4} for selective schools

In Section 6 we only used data for non-selective secondary schools. One question, therefore, is to what extent the model also works for selective schools (that is, schools where the admissions policy includes selection on the basis of ability).

(a) For the intake of some selective schools, $P_{\text{KS2}} = 31.0$. Calculate the predicted value of P_{KS4} for these schools using the least squares regression line you calculated in Activity 29.

(b) The 95% confidence interval for the predicted value you calculated in part (a) is 82.7% to 93.8%. Interpret the meaning of this interval.

(c) In 2011, the highest value of P_{KS2} for a selective school is given as 33.5. For such a school, the predicted value of P_{KS4} is 114.8%, and the 95% prediction interval is 93.4% to 136.2%. The actual value of P_{KS4} turned out to be 100%. Do the predicted value and prediction interval seem reasonable for such schools?

Exercise 14 Exploring some data about selective schools

The values of P_{KS2} and P_{KS4} for 15 selective English secondary schools are given in Table 11.

Table 11 P_{KS2} and P_{KS4} in 2011

School	P_{KS2}	P_{KS4} (%)	School	P_{KS2}	P_{KS4} (%)
1	33.4	100	9	32.8	99
2	30.9	95	10	31.8	97
3	33.0	100	11	32.2	100
4	32.6	99	12	32.7	100
5	33.5	100	13	33.1	100
6	32.2	99	14	30.9	91
7	31.3	95	15	31.4	100
8	32.1	98			

(a) Calculate the correlation coefficient between P_{KS2} and P_{KS4} using the data in Table 11.

(b) A scatterplot of the data given in Table 11 is shown in Figure 37.

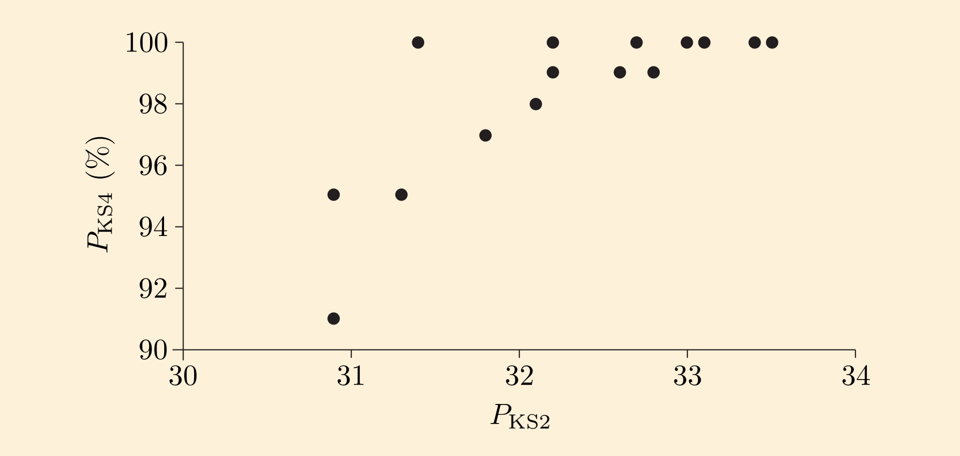


Figure 37 P_{KS2} and P_{KS4} for 15 selective secondary schools in England in 2011

Is the correlation coefficient you calculated in part (a) a good summary of the strength of the relationship between P_{KS2} and P_{KS4}? Why or why not?

Summary

In this unit the correlation coefficient was introduced – a quantity that measures the strength of a linear relationship between two variables. The correlation coefficient takes values between -1 and $+1$. When there is an exact positive relationship, the correlation coefficient takes a value of $+1$; when there is an exact negative relationship, the correlation coefficient takes a value of -1; and when there is no relationship, the correlation coefficient is zero.

You have learned how to calculate the correlation coefficient by hand and by using Minitab. When calculating the correlation coefficient, it does not matter which variable is treated as the explanatory variable, or how the variable is scaled. Outliers will tend to make the correlation coefficient appear too small, while influential points will tend to make the correlation coefficient appear too large.

You have also learned about interval estimates. In general they consist of intervals for population quantities, such as a mean or difference between means. Intervals are constructed in such a way that a known percentage of them contain the true value of the quantity of interest.

Confidence intervals for the population mean, μ, are linked to z-tests. For one sample, a 95% confidence interval for μ includes all values for which we cannot reject H_0 at the 5% significance level. When there are two samples, a 95% confidence interval includes all the values for the difference between two means that we cannot reject at the 5% significance level. 99% confidence intervals have similar definitions and are wider than the corresponding 95% intervals. You have learned how to calculate such confidence intervals by hand.

Confidence intervals for the mean response are linked to the least squares regression line. They indicate intervals for the position of the population regression line for particular values of the explanatory variable. Prediction intervals are also linked to the least squares regression line. They provide an interval for the predicted value for an individual. Prediction intervals are wider than confidence intervals for the mean response, because even if the position of the line were known, there would still be uncertainty about the value taken by an individual item. You have learned how to obtain confidence intervals for the mean response and prediction intervals using Minitab.

Finally, you have used correlation coefficients and interval estimates to explore the quality of schools – in particular, to investigate how the GCSE headline score depends on other aspects of secondary schools. Thus you have gone some way to answering the questions: *What factors influence the quality of a school?* and *How good is a school?*

Learning outcomes

After working through this unit, you should be able to:

- understand the concept of the correlation coefficient – in particular, how it relates to a relationship between two variables shown on a scatterplot
- calculate the correlation coefficient by hand
- calculate the correlation coefficient using Minitab
- roughly estimate a correlation coefficient from a scatterplot
- recognise outliers and influential points on a scatterplot, and understand their impact on the correlation coefficient
- calculate a confidence interval for a population mean
- interpret a confidence interval
- calculate a confidence interval for the difference between two population means
- interpret a confidence interval for estimating the mean response using a least squares regression line
- interpret a prediction interval for individual predictions from a least squares regression line
- obtain confidence intervals and prediction intervals for estimates from a least squares regression line using Minitab.

Solutions to activities

Solution to Activity 1

Most people would probably agree that a good school is one that provides a good education for all its students. However, what is meant by a good education is not simple to define. Enabling students to gain knowledge and skills is part of it. But you also may have thought of other aspects such as allowing students to develop artistic or sporting interests or having a positive ethos.

Solution to Activity 2

The academic ability of a student can be measured by assessing the student. This assessment might be a formal exam, a test taken in the classroom, project work, or a teacher's assessment of work done in class or of work done at home. Also, the assessment might be on a narrow range of subjects, such as English, Mathematics and Science. Or the assessment could include a much wider range of subjects

Solution to Activity 3

Using our definition of the quality of a school, the higher a school's GCSE headline figure (P_{KS4}), the better it is. However, what has not yet been defined is how high P_{KS4} needs to be for the school to be 'good'. So it is not possible to say whether this school is good or not good.

Solution to Activity 4

The histogram shows the variation between schools. For some schools P_{KS4} is more than 80%, whereas for other schools P_{KS4} is less than 30%. The median P_{KS4} is somewhere between 40% and 60%. The distribution is roughly symmetric.

Solution to Activity 5

(a) From Table 1, the number of schools with a P_{KS4} of less than 50% is $4 + 8 + 17 + 22 = 51$. So, using the criterion $P_{\text{KS4}} \geq 50\%$, 51 schools in the sample would be deemed not good enough.

(b) From Table 1, the number of schools with a P_{KS4} of less than 30% is $4 + 8 = 12$. So, using the criterion $P_{\text{KS4}} \geq 30\%$, 12 schools in the sample would be deemed not good enough.

(c) From Table 1, all but one of the schools had a value of P_{KS4} less than 90%. So, using the criterion $P_{\text{KS4}} \geq 90\%$, 99 schools in the sample would be deemed not good enough.

Solution to Activity 6

Let 'C' denote quantities relating to community schools, and let 'O' denote quantities relating to 'other' schools.

The null and alternative hypotheses are

$$H_0 : \mu_{\text{C}} = \mu_{\text{O}}$$
$$H_1 : \mu_{\text{C}} \neq \mu_{\text{O}},$$

where μ_{C} and μ_{O} are the population means of P_{KS4} of interest.

We have:

$$\overline{x}_{\mathsf{C}} = 49.8 \qquad \overline{x}_{\mathsf{O}} = 50.7 \qquad n_{\mathsf{C}} = 43 \quad n_{\mathsf{O}} = 57$$
$$s_{\mathsf{C}} = 13.55 \quad s_{\mathsf{O}} = 19.61.$$

Note that both n_{C} and n_{O} are greater than 25, so we can assume that the z-test is applicable.

The estimated standard error is

$$\mathsf{ESE} = \sqrt{\frac{s_{\mathsf{C}}^2}{n_{\mathsf{C}}} + \frac{s_{\mathsf{O}}^2}{n_{\mathsf{O}}}} = \sqrt{\frac{13.55^2}{43} + \frac{19.61^2}{57}} \simeq 3.319,$$

and

$$z = \frac{\overline{x}_{\mathsf{C}} - \overline{x}_{O}}{\mathsf{ESE}} \simeq \frac{49.8 - 50.7}{3.319} \simeq -0.27.$$

The critical values are 1.96, -1.96 (5%) and 2.58, -2.58 (1%). Since $-1.96 < -0.27 < 1.96$, we cannot reject the null hypothesis at the 5% level. There is little evidence that the mean P_{KS4} in community schools is different from its mean value in other schools.

Solution to Activity 7

In Figure 7, the relationship between the variables is weak. At best the points only loosely follow a line.

In Figure 7, the relationship between the variables is strong. It is possible to draw a straight line such that all the points lie close to the line.

The relationship in Figure 7 is difficult to classify as strong or weak. It is weaker than the relationship in Figure 7 and stronger than the relationship in Figure 7. But whether this makes the relationship strong or weak is down to individual judgement.

Solution to Activity 8

In (a), there is a strong positive relationship between the variables. The actual correlation coefficient is 0.92, but anything between 0.8 and just under 1.0 is a good guess. Remember that correlation coefficients always lie between -1 and $+1$, so you should not have guessed a number larger than 1.

The relationship in (b) is negative, but it is not as strong as in the first example; the points do not lie as close to a straight line. In this case, $r = -0.64$, but any guess between -0.8 and -0.4 would be reasonable.

In (c) there is only a weak relationship between the variables, but it is definitely positive. The actual value of the coefficient is 0.30.

Solution to Activity 9

(a) For each constituency, let x represent the attainment at Key Stage 2, and let y represent the attainment at Key Stage 4.

(i) The five initial sums are as follows:

$$\sum x = 508.0, \quad \sum y = 355.6,$$
$$\sum x^2 = 37\,008.1, \quad \sum y^2 = 18\,600.62, \quad \sum xy = 26\,030.33.$$

(ii) The sum of squared residuals of the x-values, the sum of squared residuals of the y-values and the sum of products of the residuals of the

x- and y-values are as follows:

$$\begin{aligned}\sum(x-\overline{x})^2 &= 37\,008.1 - \frac{508.0^2}{7}\\ &\simeq 37\,008.1 - 36\,866.285\,71 = 141.814\,29,\end{aligned}$$

$$\begin{aligned}\sum(y-\overline{y})^2 &= 18\,600.62 - \frac{355.6^2}{7}\\ &= 18\,600.62 - 18\,064.48 = 536.14,\end{aligned}$$

$$\begin{aligned}\sum(x-\overline{x})(y-\overline{y}) &= 26\,030.33 - \frac{508.0\times 355.6}{7}\\ &= 26\,030.33 - 25\,806.4 = 223.93.\end{aligned}$$

(iii) We can now calculate the correlation coefficient using the quantities calculated in step 2.

$$\begin{aligned}r &= \frac{\sum(x-\overline{x})(y-\overline{y})}{\sqrt{\sum(x-\overline{x})^2\times\sum(y-\overline{y})^2}}\\ &\simeq \frac{223.93}{\sqrt{141.814\,29\times 536.14}} \simeq \frac{223.93}{275.739\,58}\\ &\simeq 0.812\,11.\end{aligned}$$

So, the correlation coefficient for these data is 0.81 (rounded to two decimal places).

(b) In Figure 5 there is a clear positive relationship between the attainment at Key Stage 2 in 2006 and attainment at Key Stage 4 in 2011. Furthermore, this relationship looks reasonably linear. A correlation coefficient of $+0.81$ reflects this strong positive linear relationship.

Solution to Activity 10

The information given in the question means we can start at step 2.

$$\sum(x-\overline{x})^2 = 2329 - \frac{135^2}{8} = 2329 - 2278.125 = 50.875.$$

$$\sum(y-\overline{y})^2 = 226 - \frac{40^2}{8} = 226 - 200 = 26.$$

$$\sum(x-\overline{x})(y-\overline{y}) = 708 - \frac{135\times 40}{8} = 708 - 675 = 33.$$

Thus (using step 3) the correlation coefficient is

$$\begin{aligned}r &= \frac{\sum(x-\overline{x})(y-\overline{y})}{\sqrt{\sum(x-\overline{x})^2\times\sum(y-\overline{y})^2}}\\ &\simeq \frac{33}{\sqrt{50.875\times 26}} \simeq \frac{33}{36.3696} \simeq 0.907.\end{aligned}$$

So, the correlation coefficient is 0.91 (rounded to two decimal places).

Solution to Activity 11

(a) There appears to be no relationship between the two variables. Particular values for one variable do not appear to be associated with any particular values for the other variable.

(b) Ten of the points will provide positive contributions to the numerator of the correlation coefficient, whilst the other ten will provide negative contributions. So overall the positive and negative contributions will largely balance out, leading to a correlation coefficient close to zero. (In fact, the correlation coefficient is $r = +0.12$.)

Solution to Activity 12

The pairs in 1, 3 and 5 will all have a correlation of $+0.56$. This is because in each case the pair of variables is the same as the original pair in the question, except for some rescaling and moving of the origin.

The pair in 2 is likely to have a different correlation as the length of a car is not just a rescaling of its weight. That is, cars with similar lengths could have very different weights.

The pair in 4 is also likely to have a different correlation, as the fuel consumption is not just a rescaling of the miles per gallon. In fact, as the miles per gallon increase, fuel consumption will decrease, so the relationship between weight and fuel consumption is likely to be negative.

Solution to Activity 13

The relationship in (a) is clearly linear, and the relationship in (b) is clearly non-linear. So the correlation coefficient will provide a good indication about the strength of the relationship in (a) but not in (b). (In fact, although the relationships in both scatterplots look relatively strong, the correlation coefficient for (a) is $+0.98$, while that for (b) is only $+0.2$.)

In (c) and (d), the relationships do not look totally linear. However, the relationships are approximately linear. So the correlation coefficients will give approximate measures of the strength of the relationship in these two scatterplots. (It turns out that the correlation coefficient for (c) is $+0.93$, and for (d) it is $+0.95$. Both are quite close to $+1$, hence indicating strong positive relationships.)

Solution to Activity 14

Notice that, overall, there appears to be a positive correlation between the percentage of students achieving the benchmark at Key Stage 2 in 2006 and the benchmark at Key Stage 4 in 2011.

The points representing Newport West and Monmouth appear to be influential points. With those two points on the graph, your eye follows an upwardly sloping line that goes near them and that seems in keeping with the other points. If those two points were not there though, the line your eye followed would not slope upward at all, or it would slope up only a little.

Of all the points, the point representing Islwyn is the most outlying, as it lies away from the pattern of the other points. However, it is not as remote as the points representing Newport West and Monmouth.

The various correlation coefficients are as follows.

Points omitted	r
None	$+0.76$
Newport West	$+0.63$
Newport West and Monmouth	0.00
Islwyn	$+0.89$
Islwyn, Newport West and Monmouth	-0.42

Removing the two influential points decreases the correlation from $+0.76$ to zero, whereas removing the outlier increases the magnitude of the correlation from $+0.76$ to $+0.89$. Note that removing the two influential points and the outlier has a dramatic effect on the correlation – it is then negative. However, as this means that 37.5% of the data are then not analysed, this negative correlation has little validity.

Solution to Activity 15

No, it does not. Just because there is a positive correlation between chocolate consumption and the number of Nobel prize laureates does not mean that a country could increase its number of Nobel laureates by persuading its population to eat more chocolate, unfortunately. (This is quite apart from the issue of whether the number of Nobel prize laureates a country produces is a reasonable measure of the cleverness of its population.)

Solution to Activity 16

(a) Letting μ denote the population mean, the hypotheses are $H_0: \ \mu = 47.0$ and $H_1: \ \mu \neq 47.0$. The test statistic is

$$z = \frac{\overline{x} - A}{s/\sqrt{n}} = \frac{50.3 - 47.0}{17.19/\sqrt{100}} \simeq 1.92.$$

This is between -1.96 and 1.96, so H_0 is not rejected at the 5% significance level. There is little evidence that the population mean is not 47.0%.

(b) Letting μ denote the population mean, the hypotheses are $H_0: \ \mu = 53.0$ and $H_1: \ \mu \neq 53.0$. The test statistic is

$$z = \frac{50.3 - 53.0}{17.19/\sqrt{100}} \simeq -1.57.$$

This is between -1.96 and 1.96, so H_0 is not rejected at the 5% significance level. There is little evidence that the population mean is not 53.0%.

(c) Letting μ denote the population mean, the hypotheses are $H_0: \ \mu = 55.0$ and $H_1: \ \mu \neq 55.0$. The test statistic is

$$z = \frac{50.3 - 55.0}{17.19/\sqrt{100}} \simeq -2.73.$$

This is less than -1.96, so H_0 is rejected at the 5% significance level. There is moderate evidence that the population mean is not 55.0% – in fact, it appears to be less than this.

Solution to Activity 17

(a) $n = 50$, $\overline{x} = 15.62$ and $s = 6.44$.

We have ESE $= 6.44/\sqrt{50} \simeq 0.9108$.

A 95% confidence interval for the population mean is

$$\begin{aligned}&(15.62 - 1.96 \times 0.9108, 15.62 + 1.96 \times 0.9108)\\ &\quad \simeq (15.62 - 1.7852, 15.62 + 1.7852)\\ &\quad \simeq (13.83, 17.41).\end{aligned}$$

(b) A 99% confidence interval for the population mean is

$$\begin{aligned}&(15.62 - 2.58 \times 0.9108, 15.62 + 2.58 \times 0.9108)\\ &\quad \simeq (15.62 - 2.3499, 15.62 + 2.3499)\\ &\quad \simeq (13.27, 17.97).\end{aligned}$$

Solution to Activity 18

(a) The mean is

$$\overline{x} = \frac{\sum x}{n} = \frac{16\,946}{37} = 458.$$

The sum of the squared deviations is

$$\begin{aligned}\sum(x-\overline{x})^2 &= \sum x^2 - \frac{(\sum x)^2}{n} = 7\,762\,644 - \frac{16\,946^2}{37}\\ &\simeq 7\,762\,644 - 7\,761\,268\\ &= 1376.\end{aligned}$$

So the variance is

$$\frac{\sum(x-\overline{x})^2}{n-1} \simeq \frac{1376}{36} \simeq 38.22,$$

and the standard deviation is $s \simeq \sqrt{38.22} \simeq 6.1824$.

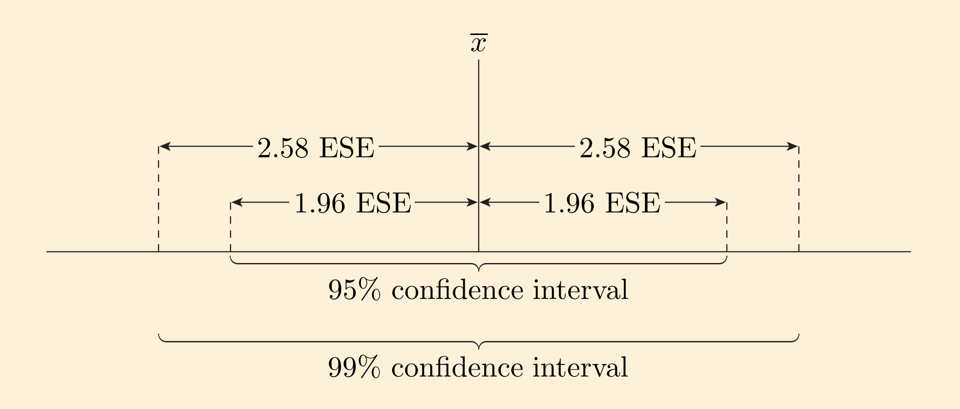
(b) ESE $\simeq s/\sqrt{n} = 6.1824/\sqrt{37} \simeq 1.016$.

Now

$$\begin{aligned}&(\overline{x} - 1.96\,\text{ESE},\ \overline{x} + 1.96\,\text{ESE})\\ &\quad\simeq (458 - 1.96 \times 1.016, 458 + 1.96 \times 1.016)\\ &\quad\simeq (458 - 1.991\,36, 458 + 1.991\,36)\\ &\quad\simeq (456.0, 460.0).\end{aligned}$$

So a 95% confidence interval for the mean weight of jars of plum jam produced by this manufacturer is $(450.0\,\text{g}, 460.0\,\text{g})$.

Solution to Activity 19



As the figure makes clear, the 99% confidence interval is always wider than the 95% confidence interval.

Solution to Activity 20

(a) From Activity 18, the 95% confidence interval for the population mean weight is $(456.0\,\text{g}, 460.0\,\text{g})$. This interval does not contain 454, so $H_0:\ \mu = 454$ is rejected at the 5% significance level. On the basis of the confidence interval, there is moderate evidence that the mean weight is not 454 grams.

(b) The confidence interval does contain 457, so $H_0:\ \mu = 457$ is not rejected at the 5% significance level. On the basis of the confidence interval, it is plausible that the mean weight is 457 grams.

Solution to Activity 21

From Example 13, the ESE for $\overline{x}_A - \overline{x}_B$ is 3.319, and $\overline{x}_A - \overline{x}_B = -0.9$, giving

$$\begin{aligned}&(-0.9 - 1.96 \times 3.319, -0.9 + 1.96 \times 3.319)\\ &\quad \simeq (-0.9 - 6.51, -0.9 + 6.51)\\ &\quad \simeq (-7.4, 5.6).\end{aligned}$$

So the 95% confidence interval for $\mu_A - \mu_B$ is $(-7.4\%, 5.6\%)$.

Solution to Activity 22

The estimated standard error is

$$\text{ESE} = \sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}} = \sqrt{\frac{18.2^2}{100} + \frac{20.6^2}{150}} \simeq 2.478.$$

Then the 95% confidence interval for $\mu_A - \mu_B$ is

$$\begin{aligned}&(\overline{x}_A - \overline{x}_B - 1.96\,\text{ESE}, \overline{x}_A - \overline{x}_B + 1.96\,\text{ESE})\\ &\quad \simeq (58.4 - 52.1 - 1.96 \times 2.478, 58.4 - 52.1 + 1.96 \times 2.478)\\ &\quad \simeq (6.3 - 4.857, 6.3 + 4.857)\\ &\quad \simeq (1.4, 11.2).\end{aligned}$$

The value 0 is not within this 95% confidence interval. Hence, the hypothesis that only children have the same average score as other children would be rejected at the 5% significance level. There is moderate evidence that the average score for only children is not the same for other children. The data suggest that in fact the average for only children is higher.

Solution to Activity 23

The equation of the population least squares line is

$$y = 16.15 + 0.48x.$$

So when $x = 6$ the population mean value for y is

$$y = 16.15 + 0.48 \times 6 = 19.03.$$

This means that the statement ‘The interval (18.31, 19.77) contains the (population) mean value of y when $x = 6$’ is true.

Solution to Activity 24

In Figure 31, seven of the 100 confidence intervals do not contain the population mean value of y (including one, sample 68, which nearly contains the population mean value), and 93 of the confidence intervals do contain the population mean value. So this means that there is a 93% chance that a randomly chosen confidence interval from these 100 intervals does include the population mean value.

Solution to Activity 25

The confidence interval means that the statement ‘The (population) mean pass rate in Mathematics is between 48.6% and 55.2% when the pass rate in English is 50%’ has a 95% chance of being true.

Solution to Activity 26

(a) The completed table is as follows.

Pass rate in English	Confidence interval for mean pass rate in Mathematics	Width of confidence interval
40	(40.1%, 49.3%)	9.2%
50	(48.6%, 55.2%)	6.6%
60	(56.9%, 61.3%)	4.4%
70	(64.5%, 68.1%)	3.6%
80	(71.0%, 75.9%)	4.9%

(b) The narrowest of the confidence intervals corresponds to a pass rate in English of 70%. A confidence interval is narrowest at the sample mean of x. Thus the average pass rate in English is 70% (to the nearest 10%).

Solution to Activity 27

Ali's prediction interval is narrower than the corresponding confidence interval, when it should be wider.

From the confidence interval for the mean response, the predicted value must be $(35.9\% + 46.4\%)/2 = 41.15\%$. However, the centre of Charlie's interval is only $(17.3\% + 55.0\%)/2 = 36.15\%$.

Solution to Activity 28

(a) The results at Key Stage 2 and Key Stage 4 have a positive linear relationship. This relationship appears to be reasonably strong.

(b) Yes, it is, because the relationship appears to be reasonably linear.

(c) There are no outliers. All the points seem to follow roughly the same trend.

Two points might be considered as influential, the two schools with the highest value of P_{KS2}. Here the performance at Key Stage 4 appears to be in line with what might be expected given the students' ability at Key Stage 2. That is, the more able a school's intake, the better those students tended to do at the end of Key Stage 4.

(d) The correlation coefficient does look reasonable. A value of $+0.83$ corresponds to a reasonably strong positive relationship – just like that noted in part (a).

Note that the correlation coefficient without the two potential influential points is $+0.81$, whereas for the full dataset it is $+0.83$. So, any impact of the potential influential points is small.

Solution to Activity 29

There are 100 observations, so $n = 100$. The slope of the least squares regression line, b, is given by the formula

$$b = \frac{\sum(x-\overline{x})(y-\overline{y})}{\sum(x-\overline{x})^2}.$$

Now

$$\begin{aligned}\sum(x-\overline{x})^2 &= \sum x^2 - \frac{\left(\sum x\right)^2}{100}\\ &= 75\,449.44 - \frac{2743.6^2}{100}\\ &= 75\,449.44 - 75\,273.4096 = 176.0304,\end{aligned}$$

and

$$\begin{aligned}\sum(x-\overline{x})(y-\overline{y}) &= \sum xy - \frac{(\sum x)\times(\sum y)}{100}\\ &= 139\,903.0 - \frac{2743.6\times 5031}{100}\\ &= 139\,903.0 - 138\,030.516 = 1872.484.\end{aligned}$$

So $b = 1872.484/176.0304 = 10.637$ (to three decimal places).

The intercept of the least squares regression line, a, is given by the formula $a = \overline{y} - b\overline{x}$.

Now,

$$\overline{x} = 2743.6/100 = 27.436 \quad \text{and} \quad \overline{y} = 5031/100 = 50.31.$$

So

$$\begin{aligned}a &\simeq 50.31 - 10.637\times 27.436\\ &\simeq 50.31 - 291.84\\ &= -241.53.\end{aligned}$$

The equation of the least squares regression line is therefore $y = -241.53 + 10.637x$.

Solution to Activity 30

(a) Between 10% and 20%.

(b) Between 55% and 60%.

(c) Between 75% and 80%.

Solution to Activity 31

From Activity 29,

$$\sum(x-\overline{x})^2 = 176.0304 \quad \text{and} \quad \sum(x-\overline{x})(y-\overline{y}) = 1872.484.$$

Also,

$$\begin{aligned}\sum(y-\overline{y})^2 &= \sum y^2 - \frac{(\sum y)^2}{100}\\ &= 282\,359 - \frac{5031^2}{100}\\ &= 282\,359 - 253\,109.61 = 29\,249.39.\end{aligned}$$

So

$$\begin{aligned}r &= \frac{\sum(x-\overline{x})(y-\overline{y})}{\sqrt{\sum(x-\overline{x})^2\times\sum(y-\overline{y})^2}}\\ &= \frac{1872.484}{\sqrt{176.0304\times 29\,249.39}}\\ &\simeq \frac{1872.484}{2269.0927}\\ &= 0.83 \text{ (rounded to two decimal places).}\end{aligned}$$

This is the same as the value given in Activity 28.

Solution to Activity 32

(a) In Activity 29 the least squares regression line was found to be

$$y = -241.53 + 10.637x,$$

where x is the Key Stage 2 average points score and y is the GCSE headline figure.

For School A, $x = 29.3$, so

$$\begin{aligned} y &= -241.53 + 10.637 \times 29.3 \\ &\simeq -241.53 + 311.66 = 70.1 \text{ (to 1 d.p.)}. \end{aligned}$$

For School B, $x = 28.1$, so

$$\begin{aligned} y &= -241.53 + 10.637 \times 28.1 \\ &\simeq -241.53 + 298.90 = 57.4 \text{ (to 1 d.p.)}. \end{aligned}$$

For School C, $x = 25.5$, so

$$\begin{aligned} y &= -241.53 + 10.637 \times 25.5 \\ &\simeq -241.53 + 271.24 = 29.7 \text{ (to 1 d.p.)}. \end{aligned}$$

So, in School A the predicted GCSE headline figure is 70.1%; in School B it is 57.46%; and in School C it is 29.7%.

(b) For School A, the residual is $63 - 70.1 = -7.1$.

For School B, the residual is $57 - 57.4 = -0.4$.

For School C, the residual is $51 - 29.7 = 21.3$.

In School A, the pupils did worse at the end of Key Stage 4 than is predicted by their results at Key Stage 2.

In School B, the pupils did about the same at the end of Key Stage 4 as their results at Key Stage 2 predict.

In School C, the pupils did better at the end of Key Stage 4 than is predicted by their results at Key Stage 2.

So, it appears that despite having the worst Key Stage 4 results, School C did the best job given the apparent ability of its intake. Similarly, even though School A had the best results at Key Stage 4, the school does not appear to have made the most of the potential of its students.

Solution to Activity 33

The actual percentage of pupils in School B achieving the Key Stage 4 benchmark is well within the prediction interval given. (As the actual and predicted percentages are so close, it would be strange if this were not the case!)

In School C the actual percentage of pupils achieving the Key Stage 4 benchmark is above the prediction interval. So there is evidence that its pupils are doing distinctly better than would be expected from their performance at Key Stage 2.

Solution to Activity 34

No, it does not. We would expect about 5% of schools to lie outside a prediction interval – that is, about five of the 100 schools plotted. So having six schools lying outside the prediction interval in Figure 36 is not much more that we would expect if the model is correct.

Solution to Activity 35

If we are not taking the Key Stage 2 results into account, then just the schools at the top of the plot (high P_{KS4}) will be classified as good, and the schools towards the bottom of the plot (low P_{KS4}) classified as bad.

One of the two schools that lie below the 95% prediction interval has a value of P_{KS4} that is sufficiently low that it might still be judged as bad on the basis of just its Key Stage 4 results.

The other five schools that lie outside the 95% prediction interval do not have a value of P_{KS4} that is noticeably either high or low. So, when the Key Stage 2 results are ignored, these schools will not be picked out as noticeably good or bad. This means, in particular, that the schools that performed well, given the ability of their intake, no longer stand out.

Solutions to exercises

Solution to Exercise 1

Factors 1 and 3 could both be tested using a two-sample z-test. This is because these factors split the schools into two groups (sixth form or no sixth form; allows early GCSEs or does not allow early GCSEs). The mean P_{KS4} could be calculated for each group, and the two means could then be compared in a hypothesis test.

The other two factors (size of school and proportion eligible for free schools meals) would have to be split into two groups before a two-sample z-test could be used. Such a split is likely to be arbitrary, making it difficult to interpret the results from the test.

Solution to Exercise 2

From strongest to weakest the order of the correlations is as follows.

$$-0.99 \quad +0.80 \quad -0.30 \quad +0.25 \quad +0.04$$

This is because the correlation gets weaker as the coefficient gets closer to zero. The sign of the correlation coefficient is unimportant.

Solution to Exercise 3

The correlation coefficient turns out to be $+0.74$. Correlation coefficients are hard to estimate accurately, so a reasonable guess would be somewhere between $+0.50$ and $+0.95$.

Solution to Exercise 4

The five initial sums that are required are as follows:

$$\sum x = 184.89, \quad \sum y = 92.06,$$
$$\sum x^2 = 3798.2743, \quad \sum y^2 = 941.7382, \quad \sum xy = 1891.2141.$$

As $n = 9$,

$$\begin{aligned}\sum(x-\overline{x})^2 &= 3798.2743 - \frac{184.89^2}{9} = 3798.2743 - 3798.2569\\ &= 0.0174,\end{aligned}$$

$$\begin{aligned}\sum(y-\overline{y})^2 &= 941.7382 - \frac{92.06^2}{9}\\ &\simeq 941.7382 - 941.6715 \text{ (rounded to four decimal places)}\\ &= 0.0667,\end{aligned}$$

and

$$\begin{aligned}\sum(x-\overline{x})(y-\overline{y}) &= 1891.2141 - \frac{184.89 \times 92.06}{9}\\ &\simeq 1891.2141 - 1891.2193 \text{ (rounded to four decimal places)}\\ &= -0.0052.\end{aligned}$$

We can now calculate the correlation coefficient.

$$\begin{aligned}r &= \frac{\sum(x-\overline{x})(y-\overline{y})}{\sqrt{\sum(x-\overline{x})^2\sum(x-\overline{x})^2}}\\ &\simeq \frac{-0.0052}{\sqrt{0.0174 \times 0.0667}} \simeq -0.15 \text{ (rounded to two decimal places).}\end{aligned}$$

Note that this correlation coefficient is considerably different from the correlation coefficient based on 100 sprinters. One explanation for this is that it is based on a relatively small sample, and a fairly biased one at that, as it consists of the best nine 200-metre sprinters.

Solution to Exercise 5

(a) There appears to be a positive relationship between the two variables. High values for one variable appear to be associated with high values for the other variable. Also, low values for one variable appear to be associated with low values for the other variable.

(b) Most of the points will provide positive contributions to the numerator of the correlation coefficient. Only three of the points will provide negative contributions to the correlation coefficient (and quite small contributions at that, as they are close to $(\overline{x}, \overline{y})$). So overall the correlation coefficient is going to be positive. (In fact the correlation coefficient is $r = +0.90$.)

Solution to Exercise 6

(a) No, the correlation coefficient would not change. It does not matter which variable is labelled as x.

(b) No, the correlation coefficient would not change. Adding a constant to a variable does not change a correlation coefficient.

(c) No, the correlation coefficient still would not change. Multiplying by a constant does not change a correlation coefficient, nor does adding a constant. Hence doing one followed by the other does not change it.

Solution to Exercise 7

The point representing Harry Aikines-Aryeety appears to be an outlier. His times lie away from the general pattern of the other points for both the 100 metres and the 200 metres. The point representing Leon Baptiste belongs to the general pattern. However, with a time of more than 10.40 seconds for the 100 metres, his time is noticeably slower than the other sprinters in this elite group. So the point representing Leon Baptiste might be counted as an influential point. If his point is removed, the correlation changes sign!

The various correlation coefficients are as follows.

Points omitted	r
None	-0.15
Harry Aikines-Aryeety	-0.62
Leon Baptiste	$+0.10$
Harry Aikines-Aryeety and Leon Baptiste	-0.46

Removing the outlier increases the magnitude of the correlation coefficient from 0.15 to 0.62, whereas removing the influential point decreases the magnitude to 0.10.

Solution to Exercise 8

(a) On the plot, the point representing pineapples appears to be to be an outlier. It is rated as being relatively tasty but very difficult. This is at odds with the positive correlation between tastiness and easiness implied by the positioning of the other fruits on the plot. (Fruits that are relatively easy are also seen as relatively tasty.)

There do not appear to be any influential points. None of the fruits are unusually tasty, unusually untasty, unusually difficult or unusually easy.

(b) No, the conclusion is not appropriate. Although easy to eat fruits also tend to be the ones that are tasty, this does not mean that being easy to eat *causes* a fruit to be tasty. (Nor does being tasty cause a fruit to be rated as being easy to eat.)

Solution to Exercise 9

(a) $$\text{ESE} = s/\sqrt{n} = 5.1/\sqrt{30} \simeq 0.931.$$

Thus, a 95% confidence interval for the mean skull breadth (in millimetres) in that period is

$$\begin{aligned}&(\overline{x} - 1.96\,\text{ESE},\ \overline{x} + 1.96\,\text{ESE})\\ &\quad\simeq (131.4 - 1.96 \times 0.931, 131.4 + 1.96 \times 0.931)\\ &\quad\simeq (131.4 - 1.825, 131.4 + 1.825)\\ &\quad\simeq (129.6, 133.2).\end{aligned}$$

(b) The 99% confidence interval would be wider. This is because a 99% confidence interval is always wider than the 95% confidence interval.

Solution to Exercise 10

The estimated standard error is

$$\text{ESE} = \sqrt{\frac{3.4^2}{30} + \frac{5.1^2}{30}} \simeq 1.119.$$

Thus the 95% confidence interval for the change in mean skull breadth (in mm) from 4000 BC to 1850 BC is

$$\begin{aligned}&(134.5 - 131.4 - 1.96 \times 1.119, 134.5 - 131.4 + 1.96 \times 1.119)\\ &\quad\simeq (3.1 - 2.193, 3.1 + 2.193)\\ &\quad\simeq (0.9, 5.3).\end{aligned}$$

As 0 mm is not in the 95% confidence interval, there is reasonably strong evidence that the mean skull breadth changed between 4000 BC and 1850 BC. With 95% confidence, the mean breadth increased by between 0.9 mm and 5.3 mm during this time.

Solution to Exercise 11

(a) When $x = 110$,

$$y = 4.2 + 0.880 \times 110 = 4.2 + 96.8 = 101.0.$$

So for patients with an initial diastolic blood pressure of 110 mmHg, the predicted post-injection blood pressure is 101.0 mmHg.

This estimate does not give any information about the uncertainty in the estimate. This information is important if the estimate is going to be used to make decisions.

(b) Yes, it does. This is because there is a high chance that it is correct to say that the mean post-injection diastolic blood pressure is in the range 95.9 mmHg to 106.1 mmHg – values that are all lower than the initial diastolic blood pressure.

(c) No, it does not. The confidence interval provides information about the uncertainty of the estimate for the population mean – not that of individuals. For that, a prediction interval is required.

In fact, the prediction interval is 81.2 mmHg to 120.8 mmHg. This interval includes values above 110 mmHg, which means that for any individual patient, their diastolic blood pressure might actually go up after treatment with captopril, not down.

Solution to Exercise 12

(a) There is a 95% chance that the statement 'the sprinter's best time for the 100-metre sprint in 2011 would have been between 9.88 seconds and 10.62 seconds' is true.

(b) The point estimate is always in the middle of the prediction interval. As $(9.88 + 10.62)/2 = 10.25$, the point estimate must have been 10.25 seconds.

(c) Yes, it is plausible, because the B qualifying time is within the range of times given by the 95% prediction interval.

(d) Yes, they do. Seven of the points lie outside of the 95% prediction intervals. This is not much more than the five that would be expected.

Solution to Exercise 13

(a) From Activity 29 we had that

$$P_{\mathsf{KS4}} = -241.53 + 10.637P_{\mathsf{KS2}}.$$

So when $P_{\mathsf{KS2}} = 31.0$, the predicted value of P_{KS4} is

$$-241.53 + 10.637 \times 31.0 = 88.2 \text{ (to 1 d.p.).}$$

Thus for a school where $P_{\mathsf{KS2}} = 31.0$, it is predicted that 88.2% of its students finish Key Stage 4 with at least five A* to C grade GCSEs including English and Mathematics.

(b) Consider the statement 'The (population) average of P_{KS4} for schools where at intake $P_{\mathsf{KS2}} = 31.0$ is between 82.7% and 93.8%.' This confidence interval means that there is a 95% chance that this statement is true.

Note that this interval refers to the average of P_{KS4} for all such schools, and not the value of P_{KS4} for any single school.

(c) No, they do not. It is not possible to get more that 100% of students passing at least five A* to C grade GSCEs including English and Mathematics. So the maximum possible value of P_{KS4} is 100%. This is less than the predicted value. Also, most of the range quoted for the prediction interval is above 100%.

Solution to Exercise 14

(a) Let y represent P_{KS4}, and x represent P_{KS2}. From Table 11:

$$\sum x = 483.9, \quad \sum y = 1473,$$
$$\sum x^2 = 15\,620.91, \quad \sum y^2 = 144\,747, \quad \sum xy = 47\,543.7.$$

So

$$\sum(x-\overline{x})^2 = 15\,620.91 - 483.9^2/15 = 10.296,$$
$$\sum(y-\overline{y})^2 = 144\,747 - 1473^2/15 = 98.4,$$
$$\sum(x-\overline{x})(y-\overline{y}) = 47\,543.7 - (483.9 \times 1473)/15 = 24.72.$$

So, the correlation coefficient, r, is as follows.

$$r = \frac{\sum(x-\overline{x})(y-\overline{y})}{\sqrt{\sum(x-\overline{x})^2 \times \sum(y-\overline{y})^2}} = \frac{24.72}{\sqrt{10.296 \times 98.4}} \simeq 0.78.$$

(b) There is some suggestion that the relationship between P_{KS2} and P_{KS4} is non-linear because P_{KS4} has to level out at 100 or below. So the correlation coefficient probably underestimates the strength of the relationship between P_{KS2} and P_{KS4}.

Acknowledgements

Grateful acknowledgement is made to the following sources:

Figure 24 Taken from: www.bbc.co.uk/news/magazine-20356613

Figure 25 Adapted from: www.topsinathletics.info

Subsection 1.1 cartoon at end (Beagle Elementary School), Rod Maclean / www.justkiddingcartoons.com

Subsection 1.1 photo (GCSE certificate), www.slideshare.net/Muppet1971/gcse-english-summer-1989

Subsection 1.2 photo (SATS papers), taken from: www.sats-past-papers.co.uk/images/S8300645.jpg

Subsection 2.2 cartoon ('Correlation Street'), www.cartoonstock.com

Exercises on Section 2, photo (Marlon Devonish) © MaxiSports / www.dreamstime.com

Subsection 3.4 cartoon (causation): Jyrinx/www.flickr.com/photos/26353267@N03/4026814703. This file is licensed under the Creative Commons Attribution-Non-commercial Licence http://creativecommons.org/licenses/by-nc/3.0/

Exercises on Section 3, figure (tastiness of fruit), www.causeweb.org

Subsection 4.2 photo (child writing): With Permission: Heather Whitaker

Subsection 4.2 photo (jam), AVRORRA / www.istockphoto.com

Subsection 4.3 photo (family) © imtmphoto / www.istockphoto.com

Exercises on Section 4, photo (Egyptian mask), taken from: www.touregypt.net/featurestories/masks.htm

Subsection 5.1 photo (Scrooge), taken from: www.stjohnec.org/portfolio/a-christmas-carol-class-of-2013-adrianne-as-scrooge

Exercises on Section 5, cartoon (sprinters), www.cartoonstock.com

Section 6 cartoon (decent school), www.cartoonstock.com

Every effort has been made to contact copyright holders. If any have been inadvertently overlooked the publishers will be pleased to make the necessary arrangements at the first opportunity.

Index